Technische Universität München
Fakultät für Informatik
Forschungs- und Lehreinheit XI
Angewandte Informatik / Kooperative Systeme

LIMITS AND CHANCES OF SOCIAL INFORMATION RETRIEVAL

CHRISTOPH FUCHS

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r):		Univ.-Prof. Dr. Dr. h.c. Javier Esparza
Prüfer der Dissertation:	1.	Priv.-Doz. Dr. Georg Groh
	2.	Univ.-Prof. Dr. Helmut Krcmar

Die Dissertation wurde am 30.03.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 05.08.2016 angenommen.

ABSTRACT

The prevailing approaches for web search are mainly driven by content similarities and disregard social relationships between the information seeker and the information provider. Furthermore, only explicitly published information is considered. Although several social search approaches exist, only a small subset interprets social search as querying other people's information spaces.

Following concepts like homophily from the social sciences, the objective of this thesis is to assess the potential of social information retrieval approaches to satisfy information needs. Therefore, a specific, but also highly customizable social information retrieval concept is developed, prototypically implemented, and evaluated in various usage scenarios. The results allow to identify limits of and success factors for social information retrieval systems.

By conducting a survey with 112 participants, we show that using one's social network is a valid method to satisfy information needs, but privacy is considered as a potential threat for information seekers (an additional survey also confirmed the results for information providers, $n = 608$). The analysis of two large social networking datasets from Twitter and Facebook indicate that content from socially close people is perceived as more important by the information seeker than content from other people, affirming social information retrieval as promising method to satisfy information needs. As part of the thesis, a social information retrieval concept is developed that is specific and specific enough to be implemented prototypically, but also sufficiently flexible and parameterizable to cover a broad range of social information retrieval scenarios. The distributed character of the system leads to smaller document collections which allow to apply semantically richer modeling approaches like latent topic models or explicit concept representations. Using these prototypes, various aspects of the social information retrieval workflow are evaluated using (1) datasets covering socially relevant information (scientific abstracts as expertise profiles, social question & answer platforms) and (2) data obtained from a real-world social information retrieval experiment using the developed prototypes with 121 participants in the course of three weeks. The social information retrieval experiment consists of a manual mode relying on human intelligence to route questions and reply to answers (considered as the hypothetical upper bound w.r.t. quality), an automatic mode (routing and content identification done by the system), and a specific use case (social product search).

The results confirm that an adjusted interaction pattern successfully mitigates the participants' reluctance to share information. The findings indicate that social closeness is positively correlated with the reply's degree of relevance. Based on the collected data, serendipitous effects can not be linked to social closeness, but appear to co-occur with high degrees of content knowledge similarity. The outcome of the social product search experiment suggests that socially close people are interested in the same products with a higher probability than socially distant people. This could be interpreted as confirmation that social networks can support buying decisions.

Overall, the results indicate that social information retrieval is a promising enhancement of existing tools for information gathering, especially for information needs that benefit from personal judgment.

ZUSAMMENFASSUNG

Vorherrschende Verfahren zur Informationssuche im Web greifen vorwiegend auf inhaltliche Kriterien zurück und ignorieren weitgehend die soziale Beziehung zwischen informationssuchendem und -bereitstellendem Nutzer. Darüber hinaus werden ausschließlich explizit publizierte Informationen berücksichtigt. Obwohl einige Social-Search-Ansätze existieren, interpretiert nur ein kleiner Teil davon "Social Search" als direkte Abfrage der Informationsräume anderer Benutzer.

Dem aus den Sozialwissenschaften entlehnten Homophilie-Begriff folgend, ist das Ziel dieser Arbeit das Potential von Social-Information-Retrieval-Ansätzen zur Erfüllung von Informationsbedürfnissen zu bewerten. Hierzu wird ein ausreichend spezifisches, aber dennoch hinreichend allgemeines Konzept eines Social-Information-Retrieval-Systems entwickelt, als Prototyp implementiert und in zahlreichen Anwendungsfällen evaluiert, um die Grenzen und Erfolgsfaktoren für Social-Information-Retrieval-Systeme zu identifizieren.

Basierend auf einer Umfrage unter 112 Teilnehmern zeigen wir, dass soziale Netzwerke eine ernstzunehmende Methode sind, Informationsbedürfnisse zu erfüllen, aber die Verletzung der Privatheit von den informationssuchenden Nutzern als potentielle Gefahr gesehen wird (eine zusätzliche Umfrage bestätigt die Ergebnisse auch für die Anbieter der Informationen, $n = 608$). Die Ergebnisse der Analyse zweier Datensätze aus dem Social-Networking-Bereich (Twitter, Facebook) deuten darauf hin, dass Inhalte von sozial nahestehenden Personen von informationssuchenden Benutzern als bedeutsamer wahrgenommen werden, was grundsätzlich Social Information Retrieval als vielversprechenden Ansatz bekräftigt. Im weiteren Verlauf der Arbeit wird ein Konzept für ein Social-Information-Retrieval-System entwickelt, das einerseits ausreichend spezifisch und konkret ist, um als Prototyp implementiert zu werden, andererseits aber auch flexibel genug ist, um eine Vielzahl möglicher Anwendungsfälle und Implementierungsvarianten abzubilden. Die verteilte Struktur des Systems und die damit einhergehende geringere Größe der einzelnen Informationsräume erlaubt die Verwendung semantisch reicher Modellierungen wie Latent Topic Models oder die Rückführung auf explizite Konzeptrepräsentationen. Mit Hilfe der Prototypen werden verschiedene Aspekte des Social-Information-Retrieval-Ablaufs evaluiert. Hierzu wird auf existierende Datensätze (wissenschaftliche Abstracts als Expertise-Profile, soziale Q&A Seiten) und empirisch erhobene Daten aus einem Social-Information-Retrieval-Experiment über drei Wochen mit 121 Teilnehmern zurückgegriffen. Das Social-Information-Retrieval-Experiment besteht aus einem manuellen Modus, der bei Routing-Entscheidungen und der Beantwortung von Fragen ausschließlich auf menschliche Intelligenz zurückgreift (um eine hypothetische, obere Qualitätsgrenze zu simulieren), einem automatischen Modus, wobei Routing von Fragen und Identifikation relevanter Inhalte durch das System durchgeführt werden, und einem konkreten Anwendungsfall (Social Product Search).

Die Ergebnisse bestätigen, dass ein angepasstes Interaktionsmuster die Teilungsbereitschaft von Informationen erhöht. Darüber hinaus weisen die Resultate darauf hin, dass soziale Nähe zwischen informationssuchendem und informationsbereitstellen-

dem Nutzer positiv mit der Relevanz des Ergebnisses korreliert. Serendipity-Effekte können anhand der gesammelten Daten nicht durch soziale Nähe erklärt werden, sondern scheinen auf Ähnlichkeiten des vorhandenen Wissens zwischen beiden Parteien zurückzuführen sein. Das Ergebnis des Social-Product-Search-Experiments bekräftigt, dass sozial nahestehende Personen mit höherer Wahrscheinlichkeit Interesse an den gleichen Produkten haben als sozial weiter entfernte Personen. Diese Erkenntnis kann als Bestätigung der Eignung sozialer Netzwerke zur Unterstützung von Kaufentscheidungen interpretiert werden.

Insgesamt lässt sich festhalten, dass Social Information Retrieval eine vielversprechende Erweiterung existierender Werkzeuge zur Sammlung von Informationen darstellt, besonders für Informationsbedürfnisse, die von persönlichen Einschätzungen profitieren.

PUBLICATIONS

Fuchs, C. and Groh, G. (2015a). An Attempt to Evaluate Chances and Limitations of Social Information Retrieval. In *Proceedings of the 5th International Conference on Advanced Collaborative Networks, Systems and Applications*, COLLA '15, St. Julians, Malta. IARIA.

Fuchs, C. and Groh, G. (2015b). Appropriateness of Search Engines, Social Networks, and Directly Approaching Friends to Satisfy Information Needs. In *Proceedings of the 5th Workshop on Social Network Analysis in Applications*, SNAA '15, Paris, France.

Fuchs, C. and Groh, G. (2016a). Data Sharing Sensitivity, Relevance, and Social Traits in Social Information Retrieval. *International Journal of Social Computing and Cyber-Physical Systems*, submitted, currently in review.

Fuchs, C. and Groh, G. (2016b). Routing of Queries in Social Information Retrieval using Latent and Explicit Semantic Cues. In *Proceedings of the Third Network Intelligence Conference*, ENIC '16, Wroclaw, Poland.

Fuchs, C., Hauffa, J., and Groh, G. (2015). Does Friendship Matter? An Analysis of Social Ties and Content Relevance in Twitter and Facebook. In *Proceedings of the Service Summit Workshop and Service Summit 2015*. Karlsruhe Institute of Technology.

Fuchs, C., Nayyar, A., Nussbaumer, R., and Groh, G. (2016a). Estimating the Dissemination of Social and Mobile Search in Categories of Information Needs Using Websites as Proxies. *arXiv:1607.02062 [cs.CY]*. `https://arxiv.org/abs/1607.02062` (retrieved 2016-11-20).

Fuchs, C., Voigt, C., Baldizan, O., and Groh, G. (2016b). Explicit and Latent Topic Representations of Information Spaces in Social Information Retrieval. In *Proceedings of the Third Network Intelligence Conference*, ENIC '16, Wroclaw, Poland.

Some ideas, results, and text fragments of this thesis have already appeared previously in the publications listed above. To improve the readability of thesis, direct quotations referencing to these publications have not been marked accordingly in the respective sections. The table in Appendix B shows in detail which parts have already appeared in which publication linked to the thesis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Search is one of the most important applications in the world wide web. With more than 140 billion searches per month in 2012[1], search engines like Google, Baidu, Yahoo!, or Bing satisfy the information needs of the majority of the internet population by identifying relevant documents matching a search query. Today's prevailing search engines rely on a global index enriched with approaches for personalization (Micarelli et al., 2007; Ghorab et al., 2013; Hannak et al., 2013) or location-awareness (Sohn et al., 2008; Church et al., 2012), conceptualization (Dong et al., 2008) (like the Google Knowledge Graph[2]), and link structure within the network (Brin and Page, 1999; Kleinberg, 1999). To find a relevant answer to an information need, the answer must have been recorded in a document, which must have been published on the web and indexed by the respective search engine. The underlying principle can be referred to as the *library paradigm* – information is codified and made available to others, who can search for it in a database (Horowitz and Kamvar, 2010; i Mansilla and de la Rosa i Esteva, 2013). The system is working well: at no single point in history, humankind had access to more information than today. Facilitated by the rapid development of technical possibilities and the need for effective tools to manage the increasing number of published documents, search approaches based on the library concept shaped the world wide web and the way we consume information profoundly.

Before writing was developed, the dominant form of information retrieval followed the *village paradigm* (Horowitz and Kamvar, 2010; i Mansilla and de la Rosa i Esteva, 2013). Information seeking was based on oral communication with the community: the information seeker at first attempted to find an answer to her question using her own knowledge – if this approach did not turn out to be successful, she defined a question and tried to identify candidates who could provide a satisfactory answer. After asking the question and receiving a reply (which could involve a forwarding step to a better suited candidate), the information seeker evaluated the response and either continued the search process (e.g., with an adapted version of the question and/or a different information provider) or learned from the response and possibly provided a reward to those who helped to answer the question. With the rise of social media, the barrier to publish own content was lowered and social relationships were explicitly modeled in online social networking platforms. Online networking platforms like Facebook publish content shared by socially close friends and follow the approach of recommending content to the users. A recent analysis[3] shows that Facebook leads more traffic to news sites than Google, the market-leading search engine provider. These findings could be interpreted as an indication for the success of social recommendation techniques. While social recommendation (Ricci et al., 2015, p. 511-543) is not the same as information retrieval (Manning et al., 2008), it shows a

1 http://www.internetlivestats.com/google-search-statistics/ (retrieved 2015-07-26)

2 https://googleblog.blogspot.de/2012/05/introducing-knowledge-graph-things-not.html (retrieved 2016-03-01)

3 http://fortune.com/2015/08/18/facebook-google/ (retrieved 2016-01-15)

lot of parallels (de Vries, 2015) and therefore supports the idea that social concepts could also positively influence information retrieval.

Even with conceptual parts borrowed from the village paradigm becoming more popular (like trust or personalization in social media or Question & Answer websites like Quora[4]), modern and automated search approaches based on the village paradigm are still in their infancy (i Mansilla and de la Rosa i Esteva, 2013). The present thesis investigates the limits and chances of social information retrieval using a prototypical implementation of a social information retrieval system. The system is designed to canonically reflect the village paradigm to stress the social element of the information retrieval process and allow a generalization of the results. The strengths and weaknesses of the search system are explored by conducting empirical experiments with a network of (social) agents who try to solve information needs by asking each other.

In the presented approach, users (automatically and manually) maintain a private information space and can query other users for information. The concept covers the integral steps of the search process for information seekers and information providers, which include (1) identifying suitable candidates to query, (2) sending queries to the information provider(s), (3) finding relevant answers in the individual private information space (performed by the information provider), and (4) replying to questions. For each of the steps, different technical implementations based on latent topic models, explicit concept representations, and privacy preserving data collection mechanisms are evaluated against each other. An ecosystem like that can not be seen as an isolated technical issue – to fully unfold its capabilities, a social search system must be designed carefully to meet human preferences in terms of interaction and information sharing. To explore the relevant social parameters of such a system, several user studies have been conducted with partial prototypes of the system. Using this setup, the thesis investigates the following research questions:

1. How do social context and interaction archetypes influence users' data sharing sensitivity in view of social information retrieval approaches?
2. Relevance and Serendipity of Results
    a) How relevant are information items taken from non-public information spaces of socially close people when satisfying information needs?
    b) Does social context imply a valuable contribution to retrieving information from the unconscious information need (serendipitous information)?
3. Which social concepts influence the users' routing decisions?
4. Which categories of information needs could benefit from social information retrieval?

A distributed social information retrieval system requires users who are willing to share information. Research Question 1 investigates how social context and the way the interaction is organized and supported by some system influence the users' willingness to share information. The first part of Research Question 2 explores how

4 `http://www.quora.com` (retrieved 2016-01-15)

far information that is hold by socially close people can be seen as a relevant source of information to satisfy one's own information need. The second part of Research Question 2 elaborates on the idea that socially close participants might possess information that fosters serendipity (i.e., satisfies the unconscious information need). Serendipity is defined as "lucky accident" – finding "interesting and inspiring information" without explicitly looking for it nor expecting to find it (Dörk et al., 2011; Dörk et al., 2012). In Research Question 3, the social attributes that are important to determine which person to query are analyzed. Research Question 4 tries to identify which types of information needs would benefit most from a distributed social information retrieval approach. Adjusting and topically or functionally limiting the approach to maximize the benefits for the narrowed-down information needs could increase the usability and acceptance of the overall solution.

The thesis follows a design science approach (Hevner et al., 2004). Therefore, the objective is to design and evaluate artifacts to increase the understanding of the problem setting and improve the proposed solution. Hevner et al. listed seven research guidelines, which shall be briefly set in relation to this work in the next paragraphs.

DESIGN AS ARTIFACT Research following the design science principle "must produce a viable artifact in the form of a construct, a model, a method, or an instantiation" (Hevner et al., 2004). As part of this thesis, several components of the proposed social information retrieval system are implemented prototypically in a way to better understand the problem domain and to gain additional knowledge about the users' perception and preferences. Examples include the prototypes used in Experiment 7 (Chapter 16, Section A.2.2) and the routing and indexing methods presented in Experiments 4, 5, and 6 (Chapter 13, Chapter 14, Chapter 15).

PROBLEM RELEVANCE The design science methodology aims "to develop technology-based solutions to important and relevant business problems" (Hevner et al., 2004). Search for information is a relevant problem. Experiments 1 and 2 (Chapter 10, Chapter 11) and previous research (i Mansilla and de la Rosa i Esteva, 2013) suggest that social means are useful sources for information and therefore a valid and relevant approach to improve information retrieval for certain types of information needs.

DESIGN EVALUATION The design science research guidelines require that "utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods" (Hevner et al., 2004). The results of the various experiments evaluating the concept are presented in each experiment's chapter. A summary of the results for all research questions is given in Chapter 17.

RESEARCH CONTRIBUTIONS According to (Hevner et al., 2004), "design-science research must provide clear and verifiable contributions in the area of the design artifact, design foundations, and/or design methodologies". The contributions of the thesis are summarized in Part IV.

RESEARCH RIGOR Hevner et al. require that design science "(...) research relies upon the application of rigorous methods in both the construction and evaluation

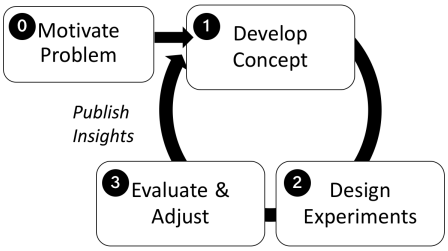


Figure 1: Research approach following the design science methodology

of the design artifact" (Hevner et al., 2004). The design artifacts have been created using and further developing state-of-the-art techniques taken from the information retrieval domain, including relatively new concepts and approaches like topic models based on LDA (Section 5.4) and explicit semantic representations based on ESA (Section 5.5). The performance of the artifacts has been evaluated with state-of-the-art tools from statistics (Section 5.2.1), including various types of regression models (linear regression, logistic regression, ordinal logistic regression) and random effects models.

DESIGN AS SEARCH PROCESS According to Hevner et al., the "search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment". Hevner interprets design science as "inherently iterative", with design being a search process "to discover an effective solution to a problem". The approach used in this thesis follows an iterative process.

COMMUNICATION OF RESULTS Hevner et al. require that "design-science research must be presented effectively both to technology-oriented (...) and management-oriented audiences". The present work can be seen as documentation of the results; some of the outcomes have already been communicated before using multiple publications listed on page vii. The thesis is structured in a way that audiences with different levels of expertise and interest in technological details can follow the overall argumentation.

Figure 1 illustrates the research approach for the present thesis: in the beginning, the problem was motivated, using Experiments 1, 2, and 3 and additional literature as proof of relevance. Based on the insights derived from the literature and the first experiments, an initial concept has been developed. The concept was evaluated with empirical experiments, each covering one or multiple parts of the social information retrieval concept. The experiments were designed to increase the understanding of the problem domain and to adjust the proposed concept.

An overview of the thesis' conceptual structure is shown in Figure 2. Part I briefly summarizes related work from the information retrieval domain, describes models for information exchange, and briefly gives some technical background on semantic web and the statistical methods used. In Part II, the proposed concept for social information retrieval is introduced. For each of the major steps in the information seeking process, a detailed concept for at least one technical solution is presented. Chapter 6 gives an overview of the complete system and the architecture. Chapter 7 describes

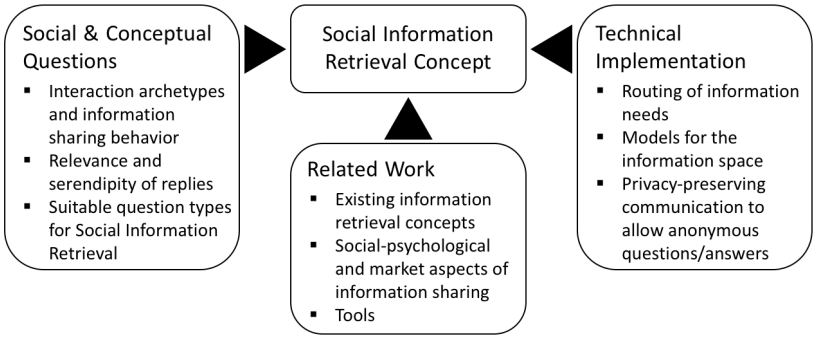


Figure 2: Conceptual structure of the thesis

the process to identify a suitable information provider, relying on the social network topology, the advertised knowledge, and a social capital market model based on previous interactions. Chapter 8 discusses multiple possibilities to organize private information spaces (which get evaluated in the empirical part of the thesis). Chapter 9 describes four exemplary scenarios where social information retrieval would be beneficial. Part III gives an overview of the empirical studies that have been conducted in the context of this thesis. While Experiments 4, 5, 6, 7b, and 7c (Chapter 13, Chapter 14, Chapter 15, Chapter 16) cover questions about the technical implementation of the system, Experiments 1, 2, 3, and 7 (Chapter 10, Chapter 11, Chapter 12, Chapter 16) explore the user's social preferences in using such a system. Part IV summarizes the results and gives an overview of the implications for social information retrieval systems. Finally, Chapter 19 concludes with an outlook on future research topics in this area.

Part I

FOUNDATIONS

In the following chapters, applied tools and related work are briefly summarized. Chapter 2 gives an introduction to Information Retrieval, covering the definition of the information retrieval process (Section 2.1), various classifications of information needs (Section 2.2), and an overview of information retrieval models (Section 2.3), context-sensitive approaches (Section 2.4), social search (Section 2.5), and privacy (Section 2.6). Chapter 3 briefly describes existing attempts to model information retrieval using market structures and foraging theories. Chapter 4 explains required concepts taken from the semantic web area, while Chapter 5 details the statistical tools and methods used in the thesis.

2

INFORMATION RETRIEVAL

Manning et al. define the term *Information Retrieval* as "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" (Manning et al., 2008, p. 1). In the following, the information retrieval process is explained (Section 2.1), a short overview of different types of information needs is given (Section 2.2), established approaches for information retrieval models are described (Section 2.3), approaches considering various forms of context in information retrieval are discussed (Section 2.4), a brief summary of social search approaches is presented (Section 2.5), and finally two examples for privacy-preserving algorithms related to the problem domain is given (Section 2.6).

2.1 INFORMATION RETRIEVAL PROCESS

According to (Hiemstra, 2009), the information retrieval process has to cover three basic steps (Figure 3):

1. Express the user's information need,
2. express the content of the documents in the document collection, and
3. match both representations.

The user translates the information need she is aware of ("conscious information need", cf. Section 2.2) to a query, i.e. a representation of the information need that can be matched against the representation of the documents in the document collection (more abstractly also referred to as information items in an information space). During the indexing phase, all documents in the collection are translated to a representation that facilitates matching with the user's query. The matching step estimates

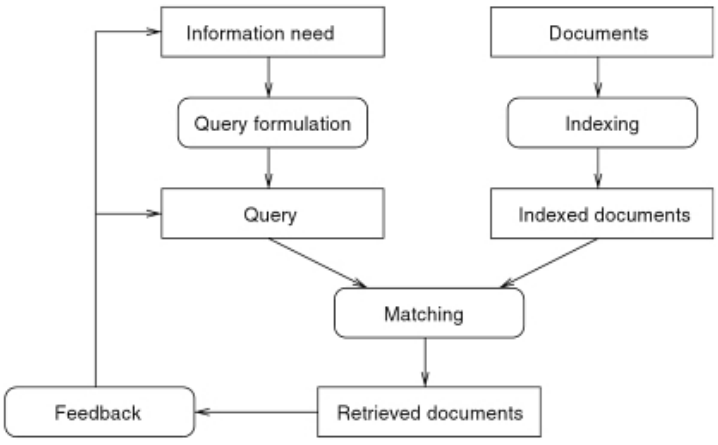


Figure 3: Information retrieval process (Göker, 2009)

whether a certain information item is "relevant" to the user's query, based on the criteria used as a relevance metric by the matching approach. The final result is a set of relevant documents (in most cases, a sorted list, depending on the specific system) where the user can choose which documents to examine further. It is likely that the user decides to adjust the information need or its representation after reviewing the result set. In certain setups, e.g. in "personalized search" (cf. Section 2.4.5), the user's implicit or explicit feedback influences the matching procedure.

2.2 INFORMATION NEEDS

2.2.1 *Models of Information Needs*

Mizzaro (Mizzaro, 1998) describes the transformation process from the original information need to an actual query as follows: The user's (objective) real information need (RIN) constitutes the objective problem the user has to solve. The user perceives the RIN as perceived information need (PIN), which is an implicit mental model of the RIN within the user's mind (i.e., a representation of a problem the user thinks she wants to solve). RIN and PIN are not necessarily equal. Using PIN as a basis, the user expresses the information need as a request, i.e., in human language. In a final step, the user translates the expression to the actual query, i.e., the input for the search system. Each of the four representations of the information need (RIN, PIN, expression, query) can describe a different set of relevant information items. This model is able to explain serendipitous results (Section 2.3.5), where information items are not necessarily considered relevant by the chosen relevance metric or the user herself, although they are relevant according to the RIN. Each transformation step on the way from the RIN to the actual query could change the set of relevant information items: it is possible that RIN and PIN do not fully overlap, e.g., because of the user's limited knowledge of the content domain. In addition, users might struggle to translate the PIN to words and the words to a meaningful query.

2.2.2 *Classifications of Information Needs*

To distinguish different types of information needs from a content perspective, several attempts have been made to define a category system for information needs. An example for an early classification schema is given in (Spink et al., 2002), where the authors examined how human information needs and search behaviors have evolved along with web content using a dataset from the Excite search engine in 1997, 1999, and 2001 (see Table 1).

Sex or Pornography	Education or Humanities
Government	People, Places or Things
Commerce, Travel, Employment or Economy	Computers or Internet
Entertainment or Recreation	Health or Sciences
Performing or Fine Arts	Society, Culture, Ethnicity or Religion
Unknown or Other	

Table 1: Classification of information needs by Spink et al. (Spink et al., 2002)

Kamvar and Baluja (Kamvar and Baluja, 2006) examined the state of web search on mobile phones in 2006 and analyzed 1,000,000 visits on Google's mobile search site, taking into account different hardware platforms (computers, keypad cell phones) to explore differences between the respective use cases. Two years later, the authors did a similar study (Kamvar et al., 2009) and compared the search patterns for searches conducted from computers, iPhones, and conventional mobile phones. The resulting categories are quite similar to the ones in the older study; however, the usage patterns changed: while in the previous study, the authors could clearly identify different patterns for search sessions conducted on a normal computer or a mobile phone, those differences were reduced for users with more powerful smartphone devices in the second study.

In (Church et al., 2007), the authors compared mobile and stationary internet usage in 2007 using a large collection of more than 30 million mobile internet requests generated by more than 600,000 European mobile subscribers over a 24-hour period in 2005. The obtained categories are listed in Table 2. Afterwards, the authors investigated information needs of mobile internet users by conducting a diary experiment (Church and Smyth, 2009) with 20 participants. The probands have been asked to document every information need that they recognize in a diary, along with additional attributes like date, time, and location. For the analysis, information needs were clustered by topics (based on the initial schema obtained in (Church et al., 2007)), by location context (away from desk, commuting, home, on-the-go, travelling abroad, work/college), and by goal (informational, geographical, and personal information management).

Adult	Multimedia
Email, Messaging & Chat	Search & Finding Things
Entertainment	Games
Unknown/Unclassified	Socializing & Dating
Shopping & eCommerce	Mobile Applications, Websites & Technologies
Sport	Auto
News & Weather	Local Services
Information	Employment

Table 2: Classification of search terms by (Church et al., 2007).

Sohn et al. (Sohn et al., 2008) investigated how mobile information needs get addressed. The term "mobile" was defined as being away from home or work. In a diary study with 20 people, the authors created a broad categorization of information needs based on the participants' diaries and their feedback. Information needs were addressed using the web (30%), calling someone who has the information (23%), calling someone who acts as a proxy to the information (16%), using external applications like Google Maps (10%), asking someone face to face (7%), referring to prepared printout (7%), going to the location (5%), and other means (2%). Other studies like (Morris et al., 2010b) cover Social Media Question Asking (SMQA, cf. Section 2.5.2) and analyzed which types of question are asked to members of one's own social network. Morris et al. (Morris et al., 2010b) conducted a survey with 624 participants about asking and answering questions on social network platforms like Facebook and Twit-

ter. The authors identified the following topic areas using an affinity diagramming technique ((Beyer and Holtzblatt, 1998), cited by (Morris et al., 2010b)): Technology, Entertainment, Professional, Restaurants, Shopping, Home & Family, Places, Current Events, Ethics & Philosophy, and Miscellaneous. Dearman et al. (Dearman et al., 2008) conducted a diary study with 20 participants to explore and analyze which information people needed and which they decided to share. The authors "(...) instructed participants to record into their diary all information they need for a task or to satisfy a curiosity, and any information they acquire throughout their everyday experiences that they would like to share with others". Their findings suggest that information needs are often "situated and contextualized"; their participants stated that contacts linked via weak ties or common contexts would be ideal candidates to provide help. These findings confirm Granovetter's theory about the strengths of weak ties (Granovetter, 1973). A more in-depth perspective on different positions on tie strength is given towards the end of Section 2.5.1. For a majority of questions, the participants stated that they would be willing to share the received information with others.

2.3 INFORMATION RETRIEVAL MODELS

In (Hiemstra, 2009) and (Manning et al., 2008), the respective authors provide a comprehensive overview of different information retrieval models. In the following, the main characteristics of each archetype are briefly summarized.

2.3.1 *Exact Match Models*

BOOLEAN MODEL In the boolean model, the query can be seen as a boolean expression that must be true for all documents that are part of the result set. A query like "university AND Munich AND informatics" would select all documents in the document collection that are indexed with all three words. A document indexed with "university", "Munich", and "computer science" would not be in the result set. The boolean model does not provide any support for ranking relevant documents because all documents fulfill the query requirement to the same extent.

REGION MODELS A region is a sequence of consecutive words within a document, defined by its start and end position. A region model works like a standard boolean model but considers text regions as default unit (instead of complete documents like the boolean model). In addition to the expressions taken from the boolean model, region models use at least the operators CONTAINING and CONTAINED_BY. Similar to the boolean model, relevance is a pure binary measure – ranking of results is not possible. To overcome this constraint, extensions have been proposed (e.g. in (Mihajlovic, 2006)).

2.3.2 *Vector Space Models*

Vector space models transform documents and queries to the same high-dimensional Euclidian vector space and calculate the similarity between the query and all available documents. Each term is represented by a dimension within the vector space. The

degree of similarity between a document vector and a query vector is interpreted as a measure for "relevance". Frequently used similarity functions include the cosine value between the vectors, the dot product (which would also consider the magnitude of the vector), or Jaccard similarity. Examples for prominent vector space models are TF-IDF (term frequency, inverse document frequency) and its enhancement, TW-IDF (term weight, inverse document frequency) (Rousseau and Vazirgiannis, 2013). In TF-IDF, each document (or query) is expressed as vector $\vec{v}$ with

$$\vec{v}_i = TF_i \cdot IDF_i = TF_i \cdot \log(\frac{N_D}{f_i}) \tag{1}$$

with TF_i representing the *Term Frequency* (how often does this specific term related to dimension i occur in the document) and IDF_i denoting the *Inverse Document Frequency*, i.e. a measure that is calculated based on the number of documents in the collection (N_D) and the number of documents that contain the respective term for dimension i (f_i). The intention is that terms that appear in many documents are not helpful to distinguish the documents among each other. TF-IDF does not reflect the position of the term inside the document (bag-of-word-assumption, (Manning et al., 2008)). In TW-IDF, the relations between words in the documents are modeled using an unweighted directed graph. Terms are represented as vertices and edges represent co-occurrences of terms within a fixed-size sliding window. The direction of the edges represents the order of the terms. With the help of this graph, meaningful term weights are extracted and replace traditional term frequencies.

2.3.3 *Probabilistic Models*

PROBABILISTIC INDEXING MODELS In Maron and Kuhns' indexing model (Maron and Kuhns, 1960), the indexer assigns each index term t a probability $P(t|d)$ given a document d. By doing this, d is not linked to t in a binary manner (yes/no) but using a more expressive probability measure. Each document is assigned to a set of index terms, weighted by their respective value of $P(t|d)$. If a user wants search for documents relevant to a specific search term, she is interested in the documents with a high value of $P(d|t)$. Using Bayes' rule, $P(d|t)$ can be rewritten as $\frac{P(t|d)P(d)}{P(t)}$. Assuming that $P(t)$ is a constant, documents can get ranked by $P(t|d)P(d)$. Here, $P(d)$ is the a-priori relevance of document d and could be defined based on usage statistics (i.e., the more often a specific document is used, the more important it seems to be). In addition, an estimate for $P(t|d)$ could be extracted by storing the search terms that have been used to retrieve d in the first place.

PROBABILISTIC RETRIEVAL MODELS The probability ranking principle, as defined in (van Rijsbergen, 1979), states that "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data."

The degree of relevance of a document d for a query q can be modeled using a random variable $R_{d,q}$, which is either 1 (document d is relevant for query q) or 0 (other-

wise). Using a probabilistic approach, ordering of results is done with $P(R_{d,q} = 1|d, q)$ (abbreviated as $P(R = 1|d, q)$ in the following). This formula can be estimated, e.g. by using the Binary Independence Model (BIM) (Manning et al., 2008): Documents and queries are represented using binary term incidence vectors, i.e. document d is represented by vector $\vec{x} = (x_1, ..., x_M)$ where $x_t = 1$ if term t is part of document d and $x_t = 0$ if it is not. It is possible that multiple documents share the same vector $\vec{x}$ (the transformation is not injective). Dependencies among terms are not considered. To model $P(R|d, q)$ using the BIM, the incidence vectors are used to represent the query and the document: $P(R|\vec{x}, \vec{q})$. Using Bayes rules, the following equations hold:

$$P(R = 1|\vec{x}, \vec{q}) = \frac{P(\vec{x}|R = 1, \vec{q})P(R = 1|\vec{q})}{P(\vec{x}|\vec{q})} \tag{2}$$

$$P(R = 0|\vec{x}, \vec{q}) = \frac{P(\vec{x}|R = 0, \vec{q})P(R = 0|\vec{q})}{P(\vec{x}|\vec{q})} \tag{3}$$

$P(\vec{x}|R = 1, \vec{q})$ reflects the probability that a relevant document for query q has the representation $\vec{x}$ (and vice versa for $P(\vec{x}|R = 0, \vec{q})$). $P(R = 1|\vec{q})$ and $P(R = 0|\vec{q})$ reflect the prior probabilities for retrieving a relevant / not relevant document for a query q. Due to the fact that a document is either relevant or not relevant for a query, the following formula holds as well: $P(R = 1|\vec{q}) + P(R = 0|\vec{q}) = 1$. For a detailed description how the probabilities can get estimated to allow ranking, please refer to (Manning et al., 2008, p. 224).

BAYESIAN NETWORK MODELS Information Retrieval using Bayesian Networks has been proposed by Turtle and Croft in (Turtle and Croft, 1990; Turtle and Croft, 1991). The main idea is to use directed graphs to model dependencies between variables. Turtle and Croft used such networks to represent information needs and documents. One part of the model is the pre-computed network of the document collection, representing the mapping between documents, terms, and concepts (derived from a thesaurus). The network representing the query needs to be computed each time a query is received and is attached to the document network. It maps from the query terms, to subexpressions of the query, to the user's information need (Manning et al., 2008).

LANGUAGE MODELS Language models are based on the idea that a document d is relevant for a query q if a probabilistic language model M_d built for the document d is likely to generate query q. Documents are therefore ranked based in their probability that their individual model created the query ($P(q|M_d)$) in decreasing order. A comprehensive introduction to language models can be found in (Manning et al., 2008, p. 237).

2.3.4 *Evaluation of Information Retrieval Systems*

A popular way to evaluate the performance of an information retrieval system on an unranked retrieval set is to assess *precision* and *recall* (Manning et al., 2008, p.

154). The elementary prerequisite is to have a dataset and a set of queries for the dataset. For each query, the (objectively) relevant documents in the dataset must be known in advance. The performance of the information retrieval system can then be measured by the ratio of (objectively) relevant documents among the documents that are considered relevant by the information retrieval system (precision) and the ratio of documents correctly considered relevant by the information retrieval system among the (objectively) relevant documents in the collection (recall).

2.3.4.1 *Precision*

The subset of documents that are (objectively) relevant to a query and are also considered as relevant by the information retrieval system are referred to as *True Positives*. The subset of documents that are identified as relevant by the information retrieval system are referred to as *Presented Elements*.

Precision is defined as

$$\text{precision} = \frac{\text{true positives}}{\text{presented elements}} \tag{4}$$

and can be interpreted as a measure the information retrieval system's ability to identify the right documents as relevant.

2.3.4.2 *Recall*

The recall value quantifies which portion of the (objectively) relevant documents are detected by the information retrieval system. It is defined as

$$\text{recall} = \frac{\text{true positives}}{\text{objectively relevant documents}} \tag{5}$$

and can be combined with the precision value explained above to calculate the widely used F1 score, defined as

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{6}$$

To apply different weights for precision and recall, the formula can get generalized. A comprehensive comparison of various metrics can be found in (Powers, 2011).

2.3.4.3 *Confusion Matrix*

Precision and recall can be inferred from the more general *confusion matrix*. Given a labeled dataset and assuming a function that predicts the relevance of a document to a query, it is possible to fill Table 3 with the respective counts for each category. The columns indicate the actual state, while the rows indicate the result of the predicting function. In the general case (i.e., with more than two classes), the error values (False Positive/False Negative) can be listed as general error counts (e.g., $E_{i,j}$ referring to the number of times where the predicted outcome belongs to row i, while the actual value belongs to column j; with $i \neq j$). As defined above, precision can be seen as the share of true positives among all items that were predicted to be relevant, while recall denotes the share of true positives among all actual relevant items (Murphy, 2012, p. 181).

		Actual	
		True	False
Predicted	True	True Positive	False Positive
	False	False Negative	True Negative

Table 3: Confusion Matrix

2.3.5 *Relevance and Serendipity*

Precision and Recall, presented in the last section, are easy to use tools to measure the performance of an information retrieval tool when using a binary scale.

However, *relevance* can be seen as a concept with multiple layers: Following Mizzaro's logic (Mizzaro, 1998) (cf. Section 2.2.1) and Groh et al.'s extension (Groh et al., 2013), where information needs can be conscious and unconscious, there is a subjective level and an objective level of relevance. Given the same real information need and two different derivations (i.e., perceived information needs), a document might be perceived as relevant by one person, but not by the other. A document might be useful for a specific person to solve a problem from the real information need given her level of knowledge, but not for the other. In addition, can a document be considered relevant if it does not explicitly answer the query, but triggers a chain of thoughts in the reader's brain which leads to an answer for the information need? A comprehensive discussion of relevance can be found in (Saracevic, 2007a; Saracevic, 2007b). The author reviews relevance from different angles and various levels of abstraction (contextual, situational, affective, cognitive, query, interface, engineering, processing, content), concluding that relevance is a timeless concept with several manifestations. With *Serendipity*, Dörk et al. describe the concept of "lucky accidents", that can – following the logic of (Mizzaro, 1998) and (Groh et al., 2013) – answer unconscious information needs (Dörk et al., 2011; Dörk et al., 2012). To formalize serendipitous information seeking behavior, Workman et al. identified the following basic properties (Workman et al., 2014):

- Serendipitous knowledge discovery (SKD) is an *iterative process*,
- SKD often involves *change or clarification* of initial information interests, which may involve *integrating new topics*,
- SKD is grounded in the user's *prior knowledge*, and
- *information organization* and *presentation* have fundamental roles.

Thudt et al. (Thudt et al., 2012) investigated how serendipitous book discoveries could be supported using information visualization and concluded that the reception of serendipity can be influenced by the following factors:

- Certain *personality traits* (e.g., curiosity)
- *Observational skills*

- *Open-mindedness* (serendipitous discoveries require "receptiveness to unexpected information"; it manifests "itself in curiosity, questioning previous assumptions, or (...) looking at information from various perspectives")
- *Knowledge*: Serendipity requires that a person is able to draw connections between seemingly unconnected information. Without prior knowledge, serendipity is not possible.
- *Perseverance*: The more time and effort one invests in a topic, the more knowledge is obtained, which in turn can foster the identification of serendipity.
- *Environmental factors*: Apart from the personal factors already mentioned, environmental factors might also impact the ability to discover and recognize serendipity.
- *Influence of people and systems*: In many cases, information is already organized by others before it is consumed. This classification can lead to serendipitous discoveries due to the fact that relations are made explicit.

Thudt et al.'s perspective on *perseverance* can be seen critically in comparison with other definitions of serendipity: they argue that obtaining more knowledge about a topic increases the chances for serendipity to happen. Following the idea of serendipity as satisfying an unconscious information need (Groh et al., 2013; Mizzaro, 1998; Dörk et al., 2011), increasing the understanding of a problem domain with additional knowledge can be seen as a shift from unconscious to conscious information needs.

Schedl et al. (Schedl et al., 2012) developed a model for serendipitous music retrieval and stated that serendipity requires similarity and dissimilarity at the same time. Bordino et al. (Bordino et al., 2013) used two datasets obtained from Wikipedia and Yahoo! Answers to investigate what makes a result serendipitous and referred to relevance and unexpectedness as components of serendipity. McCay-Peet and Toms (McCay-Peet and Toms, 2011) summarized several studies on serendipity.

Groh et al. (Groh et al., 2013) defined serendipity as "an information that is surprising to the user and has a small chance that the user might have discovered it autonomously".

2.3.6 *Architectural Classes of Information Retrieval Systems*

Traditional search engines for the web operate on publicly available documents, using a centralized global index. Distributed web search engines like (Christen, 2015) also operate on publicly available documents, but rely on a distributed global index.

Conceptually similar to traditional search engines is the concept of federated search (Callan, 2000; Govaerts et al., 2011; Lu, 2007; Arguello, 2011) where multiple collections are used to satisfy an information need. Such systems are often used as meta search engines (which forward the query to multiple other search engines and aggregate the results afterwards) or search engines that aggregate information from various sources (e.g., normal web search, image search, search in social media, etc.). A major challenge in federated search is the selection of the appropriate collection to forward the query to (Shokouhi, 2007). Baillie et al. (Baillie et al., 2011) used topic

models to support the collection selection algorithm. Comprehensive descriptions of other existing approaches can be found in (Si and Callan, 2003; Si et al., 2002); a comprehensive description of all steps required in federated search is given in (Shokouhi and Si, 2011). Federated search approaches are based on local indexes (of each collection) and are characterized by their architecture: it foresees a clear split of roles (information seeker, mediator, collections/repositories) whereas those roles are either not existent (mediator) or not defined in such a clear way (information seeker/provider) in a peer-to-peer approach (Tigelaar et al., 2012).

Approaches like the one presented in (Kontominas et al., 2013; Raftopoulou et al., 2013) use a distributed set of documents stored in individual information spaces on each client. Clients maintain a semantic index and a friend index (both stored locally) to find resources in the network. Franchi et al.'s approach called *Blogracy* (Franchi et al., 2013) consists of a peer-to-peer architecture and a local index. Documents are linked to an ontology and individual (explicit) access policies are considered. Mari et al.'s RAIS (Mari et al., 2006) consists of a distributed technical platform with local index (but a central global routing directory, the *Directory Facilitator*) to search and subscribe for information on other agents. Like (Kontominas et al., 2013), the concept shows parallels to the one presented in this thesis but also mainly focuses on technical challenges. With DIAMS (Chen et al., 2000), Chen et al. introduced an agent-based system with local indexes that allows finding and retrieving resources from other agents' information spaces. Global directories stored on *Matchmaker Agents* are used to support the routing process of the query. Information spaces are organized dynamically, a category in the information space "is both a storage for documents and an index for search and communication" (Chen et al., 2000). Categories can be nested and contain categories from other personal agents' information spaces. To index information, TF-IDF is used (cf. Section 2.3.2). The approach also includes a special feature for the search process to mitigate the problem of semantic communication gaps: "When a user query is directed to an external agent, the user's agent sends not only the query information, but also sends with the query its own query matches, i.e., its information contents related to that query" (Chen et al., 2000). An approach presented by Xu and Croft (Xu and Croft, 1999) uses clustering and language models to support distributed information retrieval. The authors compared various approaches to reorganize documents on collections (distribute documents to collections by topic; organize the documents by topic in each collection; and use relations between topic and collection as indexing function only, without reorganizing the collection by topics). The authors focused on the information's content and – in comparison to the present work – have not included any user-related theories in their concept. Tigelaar et al. assembled a comprehensive overview of other approaches for distributed information retrieval and remaining challenges (Tigelaar et al., 2012).

2.4 CONTEXT

2.4.1 *General*

Groh (Groh, 2011, p. 28) portrays the evolution of the term "context" in the domain of context-aware applications and concludes that the definition given by Dey (Dey, 2001) "may still be considered as a working consensus":

> "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves." (Dey, 2001)

The respective situation of an entity can be described in several ways. According to (Chen and Kotz, 2005), Schilit et al. (Schilit et al., 1994) identify three different basic categories of context elements with regard to mobile, ubiquitous, and wearable computing:

- *Computing context*, including network connectivity, communication attributes (costs, bandwidth, latency), or available resources (e.g. display)
- *User context*, including a user profile, location, or people nearby
- *Physical context*, including light, noise levels, traffic conditions, and temperature (Chen and Kotz, 2005)

Chen and Kotz (Chen and Kotz, 2005) propose to add *time* (e.g., "time of a day, week, month, and season of the year") as additional dimension of context.

As most other activities, searching for information is influenced by the embedding context. The relation to time is e.g. mentioned in (Mizzaro, 1998), while the authors of (Han et al., 2013) use an ontology-based approach to classify different context types (time, space (=spatial context), content) to improve the performance of a search system. The authors of (Chen et al., 2005) propose an ontology to reflect context in pervasive applications, e.g., "information about a location, its environmental attributes (e.g., noise level, light intensity, temperature, and motion) and the people, devices, objects, and software agents that it contains. Context may also include system capabilities, services offered and sought, the activities and tasks in which people and computing entities are engaged, and their situational roles, beliefs, and intentions". A similar approach is presented in (Wang et al., 2004).

In the following, related work for the most tangible context types (social, spatial, temporal, personal) will be briefly summarized.

2.4.2 *Social Context*

Groh (Groh, 2011, p. 33) distinguishes *short-term*, *medium-term*, and *long-term social context* for a user. *Short-term social context* reflects the current social situation a user is in, including the set of persons the user interacts with, a spatial reference where the situation takes place, and – as mentioned for social situations in (Groh et al., 2011b) –

a content description and a temporal reference. For higher-level representations, the emotional state could also be considered. Typically, short-term social contexts have a validity of seconds to a few hours. *Medium-term social context* "refers to time intervals from several hours to several days and encompasses derived characterizations of longer lasting social behaviors, rhythms, cohesions or events such as a visit to a friend in a foreign city, a vacation trip with others, or general interaction patterns with other people" (Groh, 2011, p. 33). *Long-term social context* refers to time-intervals that cover a much longer period, ranging from weeks to months or even years. Those relationships are typically explicitly stated as "friendships" on social networking platforms (Groh, 2011).

Social Correlation Theory is one of the most important social theories and suggests that people's behavior depends on their individual social groups and the interactions with others (Tang et al., 2014). The theory consists of three main components:

- *Homophily* refers to the fact that the probability of contact and interaction is higher for people who are similar (Watts et al., 2002). This provides an explanation why people with an overlap of interests or other attributes connect to each other easily. By implication, it also builds the logical foundation why socially close people could have relevant information in their information spaces. A prominent saying that depicts the concept of homophily is "birds of a feather flock together" (Tang et al., 2014).

- *Influence* describes the idea that users within a social environment tend to follow their socially close neighbors, who usually show similar social behaviors.

- *Cofounding* suggests that social interactions are also influenced by external factors defined by the environment that embeds the interactions.

The authors of (Tang et al., 2014) confirm that the homophily principle is valid for online social networks using datasets obtained from Twitter and Facebook.

Despite building a set of overlapping clusters with similar people, social networks also allow information transmission: In the 1960s, Stanley Milgram tried to answer the question how likely it is that two random individuals know each other. Later, he specified the question to how many intermediate individuals are needed to connect two randomly chosen persons. This question led to the *Small World Experiment* (Milgram, 1967), where random participants have been chosen from Boston, Massachusetts and Omaha, Nebraska. The participants have been asked to direct a letter to a target person in Boston by forwarding it to a friend who might be closer to the target person. The friends should do the same until the letter arrives at the target person in Boston (Tang et al., 2014; Watts et al., 2002). The result showed that the average path length of the successfully delivered letters was about six. This led to the *Six Degrees of Separation* idea, stating that on average, each person is separated from each other person on the world by about six acquaintances. The experiment did not only show that it is possible to route information efficiently in social networks, but also highlighted that it is possible to do so with local knowledge only. Since then, several publications have followed, aiming to understand and improve the routing mechanisms and providing a strong basis for social referral as a concept (Kleinberg, 2006a; Kleinberg, 2000; Kleinberg, 2006b). Although the small world experiment has

fostered a remarkable body of new research topics for sociological and behavioral studies, it is worth mentioning that only 20% of the referral chains were successful. An additional example related to information retrieval is Granovetter's concept of *weak ties* (Granovetter, 1973), stating that more distant social relations can help to access information that is more valuable than information received from *strong ties* – one of the reasons could be that socially close contacts often have access to the same sources of information as the information seeker herself.

With the importance of social networks, the interest in examples for social networks increased. In addition to explicit representations derived from online social network platforms like Facebook, Twitter, or VK.com, several studies used sensors to implicitly infer social relations. Examples include measuring physical proximity using bluetooth (Eagle and Pentland, 2006; Madan et al., 2011), radio and infrared sensors (Groh et al., 2010; Olguin et al., 2009; Madan et al., 2011), audio recording (Groh et al., 2011c), or GPS and cell tower logs (Aharony et al., 2011). In addition, social interaction can get assessed directly via call logs (Pan et al., 2011; Madan et al., 2011; Aharony et al., 2011) or contact lists (Aharony et al., 2011).

Alan P. Fiske suggested in (Fiske, 1992) that human social life could be explained by combining four psychological forms, namely *communal sharing* (CS), *authority ranking* (AR), *equality matching* (EM), and *market pricing* (MP). Following this approach, information sharing could be considered as a social act, allowing to express the underlying motivation as a combination of Fiske's forms. In CS, people treat members of their specific group as equivalents. People within the group behave altruistic towards each other and are sometimes linked by kinship. In AR, people are ordered linearly according to some social hierarchical dimension. People with higher ranks typically have privileges, prestige, and prerogatives, which people with lower ranks do not have. EM describes a relation between two people who try to keep the balance of their relationship even. This is the standard behavior among people who meet regularly and follow a tit-for-tat strategy or some other reciprocal granting of favors. In contrast, in MP relationships all relevant features are reduced to a lower dimensional value or utility metric (e.g., price) that is used to compare different factors. This is the default relationship for people who only meet once and do not plan any further encounters.

2.4.3 *Mobile Context*

In addition to social context, the users' mobile context also influences the information seekers' behavior. Mobile context has been defined in various ways: (Church and Oliver, 2011) refer to it as "being *on-the-move*", (Sohn et al., 2008) defined it as "being from home or work" (i.e., being away from places where the users spend most of their time). Other studies like (Kamvar and Baluja, 2006; Kamvar et al., 2009) only rely on the fact that a mobile interface was used. For the present thesis, we stick to the generalized definition given by Sohn et al.: A user is in a mobile if she is away from places where she spends most of her time. In comparison with non-mobile settings, mobile contexts often offer limited degrees of freedom to accomplish tasks (e.g., due to resource restrictions like screen size, (Kamvar et al., 2009)).

Church and Oliver analyzed mobile information needs of 18 participants in a 4-week online diary study (Church and Oliver, 2011). Their results suggest that contexts like location, time, activity, and social interactions have a stronger influence on the information need in mobile settings than in stationary settings. In addition, they found that "mobile search is used in more random situations and in particular is dictated by user interactions and conversations, thus highlighting a trend towards social mobile search". Church and Oliver use the term *Social Search* in a different way than it is understood in this thesis (for details on different definitions please refer to Section 2.5).

Kamvar and Baluja analyzed search patterns on Google's mobile search interface (Kamvar and Baluja, 2006) from 2005. The authors concluded that the diversity of queries is lower than on the desktop search engine, despite the fact that other attributes like words or characters per query stay the same. In addition, they noticed that searching on mobile devices takes much more time than searching on normal computers. These results need to be interpreted in the context of their origin – it is quite likely that mobile devices have improved significantly in comparison to 2005. In 2009, Kamvar et al. published another study (Kamvar et al., 2009), analyzing search logs from 2008. By then, the first Apple iPhone was available and was analyzed separately by the authors. Queries issued from normal computers or iPhones showed a higher diversity than queries from other mobile devices. Especially when comparing the results with the previous study, the hypothesis that the diversity gap between normal computers and high-end smartphones like the iPhone is shrinking can be confirmed. The authors were surprised that they could not show an increase of location-dependent queries on the iPhone in comparison to the queries issued from a stationary computer and supposed that additional applications on the iPhone running outside of the browser (e.g., Google Maps) could have caused this anomaly.

Sohn et al. (Sohn et al., 2008) conducted a diary study of mobile information needs and concluded that 75% of the mobile information needs were prompted by contextual factors like time (when need occurred), current activity, current location, and conversations taking place with others.

2.4.4 *Temporal Context*

Kramár and Bilevic investigated the temporal context of information needs using diverse hierarchical clustering (Kramár and Bilevic, 2014). The authors distinguished short-term (couple of hours to several months) and long-term interests (which could last for many years). The authors mentioned two main issues caused by the dynamics of interests: *bursts* occur when users gain a new interest and consume as much information as possible to build a solid knowledge foundation for that interest, while *drifting* describes the situation when users change their interest several times during a short period of time (Kramár and Bilevic, 2014). Their approach proposes to extract keywords from visited websites and to include additional metadata (like duration of the visit and navigation route) in the model.

2.4.5 *Personalized Search*

Personalization of web search describes an approach to improve the performance of the information retrieval system by adjusting the parameters of the search process individually to each information seeker. Therefore, an individual user model is generated to learn each user's preferences. This model can be leveraged to create a personalized ranking of the results, to allow a personalized expansion of queries, or even to conduct personalized indexing of documents. In (Micarelli et al., 2007), a comprehensive overview of different approaches is given. The authors listed current context (open documents, emails, web pages), search history (past queries, browsed websites, selected results), rich user models (user feedback on results, past queries), collaborative approaches (past queries, selected results, user ratings), result clustering (selected clusters in taxonomies), and hypertext data (selected websites, queries) as potential sources to build individual user models. In (Steichen et al., 2012), the authors compared personalized information retrieval models with adaptive hypermedia techniques. While personalized retrieval models employ a user profile as a simplified "persona" of the user, adaptive hypermedia tries to follow a multi-faceted approach consisting of many dimensions, including user goals or prior knowledge.

Ghorab et al. (Ghorab et al., 2013) also provided an extensive review of different approaches to personalize information retrieval systems, including personalization for groups (like in (Teevan et al., 2009)), multilingual systems, and social information. Recurring themes cover re-ranking results of traditional search engines, filtering results from traditional search engines, or using a scoring function within the normal IR mechanism. While the first two classes work on top of a normal IR system, the third class incorporates the personalization component in its core ranking function. Examples for re-ranking results include (Vallet et al., 2010) and (Noll and Meinel, 2007), where tags from the social bookmarking service Delicious[1] are used to annotate and re-rank URLs in the result list based on the similarity of the individual user profile (built from the tags the user has used) and the respective URL. (Micarelli and Sciarrone, 2004) explains an approach where results are filtered (i.e., removed from the result list) in accordance with the user profile (which is based on semantic networks and stereotypes to model the users' information needs). An example which includes the personalization mechanism in the IR core functionality is the approach described in (Agichtein et al., 2006), which builds relevance models considering implicit feedback from search logs. (Haveliwala, 2002) generalizes the idea of PageRank (Brin and Page, 1999) and calculates individual PageRank numbers for different topics (based on a given topic hierarchy). Qiu and Cho (Qiu and Cho, 2006) extend this approach and include individual user preferences in the PageRank calculation.

Hannak et al. (Hannak et al., 2013) developed a methodology to measure the degree of personalization on web searches by running a controlled experiment (sending predefined queries to Google from various user accounts that have small but fixed differences in their attributes and behaviors). The authors concluded that 11.7% of the search results show differences due to personalization. Shapira and Zabar proposed an approach (Shapira and Zabar, 2011) that merges recommendation and personalized information retrieval. The authors argumented that one essential reason why

1 `https://delicious.com/` (retrieved 2016-02-25)

search engines do not only provide relevant results is the fact that users provide insufficient queries (2-3 words only). Therefore, search engines try to infer additional information about the user and her information need. In addition to implicit cues like search history or contextual indicators as location, methods from the recommender systems domain measuring the similarity between users and/or queries have been proven to be useful to personalize search engines. Their approach uses social tie strength as a predictor for relevance of documents and outperforms the baseline defined by Apache Lucene[2].

Other approaches rely on the user's social network for personalization and therefore already could be seen as a personalized variant of social search, presented in the next section. In (Lu and Li, 2011), the authors used judgments from friends of the information seeker to predict her preference on returned photos. Gou et al. used the social network of the information seeker as additional input to a ranking algorithm based on TF-IDF (Gou et al., 2010) and demonstrated with a dataset extracted from YouTube that this improves the performance. The authors assumed that each searcher has one primary interest when searching (Gou et al., 2010, p. 317).

2.5 SOCIAL SEARCH

In (McDonnell and Shiri, 2011) the authors provided a classification of social search in the web domain. Following their logic, "social search" can be categorized along the following dimensions:

- *Collaboration*: synchronous vs. asynchronous,
- *Collaboration*: implicit vs. explicit,
- *Search target*: finding people vs. finding resources,
- *Search results*: Sense-making vs. content selection,
- *Finding*: Search vs. discovery

McDonnell and Shiri (McDonnell and Shiri, 2011) also categorized a comprehensive set of published social search approaches according to the dimensions above. In the following, the dimensions will be explained briefly.

COLLABORATION Collaboration among users can take place in several ways. An important factor for differentiation is whether the users need to act synchronously or not. Examples for synchronous collaboration are "joint search" and "coordinated search" described by Morris in (Morris, 2007): in the "joint search" approach, a small group of two to four users use a single computer to discuss and perform the web search jointly, while in "coordinated search" every user works on an own computer, but the team is sitting on adjacent tables so that communication is still possible (comparing of results, competing, looking at the screen of others, discussing). Motivated by the identified need of an appropriate support for this kind of user behavior, Morris and Horvitz developed a search client, SearchTogether, to foster collaboration among

2 `https://lucene.apache.org/core/` (retrieved 2015-11-20)

users while performing web searches (Morris and Horvitz, 2007). The tool allows users to form groups, enables them to see the queries used by others and the websites that have been identified as relevant. In addition, members of a group can share their search history and publish and discuss their search ideas and strategies. Social search approaches with asynchronous collaboration do not require the users to interact in real time. Examples include "using data from social media systems to improve web search" like social tagging or bookmarking. Prominent examples are the social bookmarking service Delicious[3] or Google's +1 feature[4].

Apart from interaction type, collaboration can also be categorized by whether it takes place implicitly or explicitly. Implicit collaboration happens when the involved users do not know that they are helping each other. This could occur when the activities (e.g. bookmarking a website) of one user are used to support other users (e.g., by adjusting web search ranking). Examples for this category include TC-SocialRank (Gulli et al., 2009), which considers the importance of users in their social community when evaluating the importance of the shared resources, or (Schmidt et al., 2009), where web sites are enhanced with the tags for these web sites taken from a social bookmarking service. An example for explicit collaboration is HeyStaks (Smyth et al., 2012). Users can create (and share) *staks* for search topics. The staks get filled while the user is searching. Search histories are saved anonymously in staks. When a stak member conducts a new search, the information in the stak is used to enrich the web search results, based on recent web searches performed by other stak members. By design, approaches involving synchronous collaboration are in most cases also examples for explicit collaboration.

SEARCH TARGET Some scholars characterize "social search" as finding people, e.g. Evans et al. define it as "the way individuals make use of peers and other available social resources during search tasks" (Evans et al., 2009, p. 3378). This approach shows parallels to the problems discussed in the expert search literature (cf. Section 2.5.1). According to McDonnell and Shiri, "most perspectives on social search focus on the user's desire to find information sources as the primary goal: the products of social collaboration are merely a means to obtaining the desired information" (McDonnell and Shiri, 2011, p. 12).

SEARCH RESULTS In (McDonnell and Shiri, 2011), the authors differentiate between sense-making and content selection. While a search system tries to provide links that are most relevant, the users have to decide whether a link is worth following (and reading) or not (i.e., continue the screening process of other websites). To support the users' decision, search systems typically provide information about the search results (this is similar to the concept of information scent proposed by Chi et al., (Chi et al., 2001)). This dimension differentiates whether the search system "relies on social media to either select content (i.e. provide relevancy ranking) or to help the user make sense of the search results" (McDonnell and Shiri, 2011, p. 14). Using social media to assess the relevancy of a result item could be achieved using social media

3 `https://delicious.com` (retrieved 2015-09-16)

4 `http://www.google.com/intl/en/+/learnmore/+1/` (retrieved 2015-09-16)

elements for personalization (Section 2.4.5), while sense-making could be supported by including social media content to improve the representation of the result items.

FINDING Typical approaches to social search address the classical information retrieval model of search: users have a mental representation of their information need, transform it to a query, send the query to the search system that returns a list of pointers (URLs) and summaries for each item. Ordering of result items can be performed in various ways, e.g. based on social media content or previous search sessions. Some approaches to social search do not follow the classical "search" concept but focus on the discovery of new information items. An example of the usefulness of this approach is shown in an analysis conducted by Heymann et al. (Heymann et al., 2008), which shows that 25% of the URLs posted to the social bookmarking service "delicious"[5] were new sites, not listed in a search engine. Leveraging other users' behavior to improve online navigation is referred to as "social navigation". Also other studies confirm that social navigation can help to satisfy the users' information needs (Vuorikari and Koper, 2009; Millen et al., 2007). In (Tang et al., 2012), the authors' results suggest that friends are a better source for book recommendations than people with similar reading preferences or recommendations solely based on authorship. Based on a dataset created in a study in the educational domain, Hsiao et al. reported that social navigation helps weaker students to identify relevant information (Hsiao et al., 2013).

2.5.1 *Search for Experts*

Expertise and knowledge in an organization can be a huge competitive advantage (Neef et al., 1998). Therefore, finding experts is a heavily discussed topic in the knowledge management literature and has lead to "expert finders" (Yimam, 1996) or "expertise-locator systems" (Becerra-Fernandez, 2006), a special class of search engines dedicated to finding experts within a group or organization for a given topic. Those systems have in common that a user wants to satisfy a certain "expertise need". The response of such a system is a set of experts who have a certain expertise with regard to the topic the user was searching for. Many expertise-location systems have been proposed, examples include "Who knows" (Streeter and Lochbaum, 1988), which identifies expertise using latent semantic indexing of project reports or Balog and de Rijke's approach, which relies on the company's intranet to generate expertise profiles (Balog and de Rijke, 2007). Other approaches use authored documents (Serdyukov and Hiemstra, 2008) or browsing histories (Li and Chang, 2007) to build expertise profiles for users. Zhang et al. (Zhang et al., 2011) relied on the user's search behavior to identify the user's domain knowledge: in an experiment, they recognized that the number of documents saved, the average length of a query, and the average ranking position of documents that were opened are the three variables that predict domain knowledge best. The relation between background knowledge and search has also been discussed in (Duggan and Payne, 2008; Eickhoff et al., 2014). Foner proposed Yenta (Foner, 1997), a system that identifies experts by searching their email archives in a distributed way. Zhang and Ackermann (Zhang and Ackerman, 2005)

5 `http://www.delicious.com` (retrieved 2016-01-17)

evaluated different strategies using the Enron email corpus (Klimt and Yang, 2004). One of their findings is the confirmation that weak ties play an important role in expertise identification. A prominent example of systems that rely on social referral is Katz et al.'s ReferralWeb (Kautz et al., 1997a).

Borgatti and Cross (Borgatti and Cross, 2003) proposed a conceptual framework to model the probability whether a person will be considered as information provider in a social information retrieval situation or not. Their model consists of four components, "(1) knowing what that person knows; (2) valuing what that person knows; (3) being able to gain timely access to that person's thinking; and (4) perceiving that seeking information from that person would not be too costly" (Borgatti and Cross, 2003).

Other examples integrate social aspects into their approach, e.g. Smirnova and Balog (Smirnova and Balog, 2011) did not try to find the best available expert from a content perspective. Their solution considers user-oriented factors like time to contact the expert, expected value of the knowledge gained after contacting the expert, physical distance and distance in the social graph, hierarchy, and previous collaboration. Their evaluation using a university expert search system, expert judgments and a purely content-based algorithm as a baseline revealed that physical distance and previous collaboration (co-authorship) perform better than other approaches. With *SocLaKE*, Kukla et al. leverage the underlying social network between the expertise seeker and the expert: it is more likely that an expert reacts upon a request when the request came from a person who is socially close to the expert. Therefore, Kukla et al.'s system does not recommend the actual expert but socially close friends/colleagues of the expert (Kukla et al., 2012).

Joung et al. conducted a comprehensive comparison of different strategies to find suitable experts in a social network using a large dataset extracted from an online community (Joung et al., 2013). A strategy could involve multiple intermediate steps between the expertise seeker and the expert. The authors defined a strategy as the process of identifying the right friend to forward the request to. The three main strategy clusters that have been evaluated are profile-based (PB) strategies, structural-based (SB) strategies, and hybrid approaches. Profile-based strategies work solely based on the profile of potential candidates, considering information like occupation, interest, location, gender, or marital status. SB strategies include best connection (SB-BC), hamming distance (SB-HD), or tie strength (SB-WT/SB-ST). Best connection selects the friend with the most number of friends as the recipient, hamming distance selects the one with the most number of uncommon friends, and strategies based on tie strength selects either the friend with the weakest (SB-WT) or the strongest tie (SB-ST). Tie strength is calculated based on the number of friends the expertise seeker and the potential candidate have in common. Hybrid strategies include a combination of SB and and PB, executed sequentially. The authors decided not to consider knowledge (as it might change often and fast over time). Also, profile and query similarity are only measured using term similarity (i.e., no modeling of higher level concepts has been used). In addition, the authors reported that profiles have not been maintained very well. On their test dataset, the hybrid strategy SB-PB outperforms the other strategies.

Lappas et al. summarized different approaches for expert finding in social networks (Lappas et al., 2011). They distinguished between approaches without graph elements and approaches relying on graphs to allow propagation of expertise. While solutions in the first group are normally based in information retrieval models (which model the probability $P(x|Q)$ that a candidate x is expert on topic Q using a generative model, e.g. $P(Q|x) \cdot P(x)$), the latter approaches are based on a social network, where individual expertise scores are boosted by social closeness. Examples for algorithms based on the social network structure are PageRank (Brin and Page, 1999) and HITS (Kleinberg, 1999). In addition, the authors describe approaches to form expert teams based on various social metrics.

Balog et al. also provided a large overview on different challenges and approaches to expertise retrieval (Balog et al., 2012).

Regarding the importance of tie strength in social networks for information retrieval (Granovetter, 1973), various studies came to different conclusions: while e.g. Levin and Cross (Levin and Cross, 2004), Constant et al. (Constant et al., 1996), Granovetter (Granovetter, 1983; Granovetter, 1973), Zhang and Ackerman (Zhang and Ackerman, 2005), and Rogers (Rogers, 1983) identified weak ties as more relevant for information search in a social network, other studies suggested that strong ties are more important for information exchange (Panovich et al., 2012; Hansen, 1999; Uzzi, 1997; Szulanski, 1996).

2.5.2 *Search in Social Media*

Searching social media content is different from searching general document collections (especially when considering the context of the present thesis), as it already explicitly covers social elements like relationships, trust, or reputation. To define social media, Kietzmann et al. proposed a framework with seven building blocks, namely identity, conversations, sharing, presence, relationships, reputation, and groups (Kietzmann et al., 2011). According to the authors, not all parts of the framework have to be present in a given social media scenario, they merely offer a way to discuss different archetypes on a more abstract level. The first building block, *identity*, describes how users reveal their identities in a social media scenario. This could be done by creating profiles, including personal information like name, age, gender, occupation, location, etc. but can also happen by "self-disclosure" of subjective information like thoughts or feelings. With the increasing number of social media applications, large social networking platforms like Facebook or Google Plus have been extended to provide identity services for other applications. Protocols like OAuth[6] allow to integrate various services without unveiling the user's password to a third-party service. The *conversations* building block describes to which extent users are communicating with each other in the social media scenario. Means and purpose of communication differ among services: while messages published on micro-blogging services like Twitter are of limited length, are centered around real-time status updates, and do not require answers, blog posts on the other end of the spectrum are more suited for in-depth discussion of often lengthy conversations (and not so much about staying in touch). *Sharing* defines how users "exchange, distribute, and receive content" (Kietzmann

6 `https://tools.ietf.org/html/rfc5849` (retrieved 2016-01-16)

et al., 2011). Next to conversations, sharing is the glue that connects the acting users in the social graph. A social media service needs to exploit common interests among the users to foster sharing activities to strengthen social ties. *Presence* describes to which extent users know about the current location of others (both, in the virtual and the real world) and whether the other users are available, often indicated using a status flag. Presence attributes are crucial when users want to interact in real time. *Relationships* model the social graph, and are made explicit using social relations like friendship, collaboration, or other forms of common attributes. *Reputation* describes "to which [extent] the user can identify the standing of others" (Kietzmann et al., 2011). Reputation can be seen as an indicator for trust and quality. The *Groups* building block describes to which extent it is possible for the users to organize themselves into groups and sub-groups. With the increasing number of users on social media platforms and the growing number of relations, platforms like Facebook or Twitter have introduced features to cluster friends into groups to simplify the management of social relationships. Groups also act as the surrogate for offline clubs, where members have certain additional rights (e.g., consume group content and share content with the group).

In (Zhu et al., 2013), the authors described an approach to build a topic hierarchy of user-generated content for a specific root topic and assign user-generated content from various sources to the hierarchy. In addition, it is possible to update the hierarchy once new content is available. The hierarchy could be used for (explorative) browsing or search (i.e., find the respective node and continue to browse its neighbors). Nagpal et al. presented a tool called SLANT, which uses the information in personal email and twitter feeds to augment regular web search with social content (Nagpal et al., 2012). Carmel et al. (Carmel et al., 2009) compared several social networks (familiarity network, similarity network, topic-based network) to personalize the ranking when searching for social media. Their findings suggest that the best results are obtained when relying on the combined network defined by social interactions with people who did similar things (e.g., co-posts in certain blogs, i.e. the similarity network) and the similarity of content-profiles based on individual interest profiles. Agichtein et al. proposed an approach to identify high-quality content in social media, using a dataset from a Q&A website as an example (Agichtein et al., 2008). Bhattacharyya and Wu (Bhattacharyya and Wu, 2014) introduced InfoSearch, a search engine on top of the online social networking platform Facebook, to search the content that has been shared by others.

2.5.3 *Search in Peer-to-Peer Systems*

Following the village paradigm (i Mansilla and de la Rosa i Esteva, 2013) and interpreting "social search" as asking others for help, the architectural similarities to peer-to-peer systems are obvious: peer-to-Peer systems are systems where single nodes are connected to a network and directly communicate with the other members of the network. All nodes in the network are equally privileged, each node can offer and consume services provided by other peers. The network can get clustered in logical layers to perform tasks, also called overlay-networks (Tigelaar et al., 2012, p. 9:2). Tigelaar et al. provided a wide overview on peer-to-peer information retrieval

that also covers applications, challenges, tasks and architectural archetypes (Tigelaar et al., 2012). The authors clustered peer-to-peer information retrieval systems in two categories:

- Peer-to-Peer information retrieval systems with *internal document* references, where the documents are stored on the nodes and need to get retrieved first and
- Systems with *external document* references where the actual documents are stored outside of the peer-to-peer system, e.g. a peer-to-peer based web search system.

Tigelaar et al. (Tigelaar et al., 2012, p. 9:22) gave a comprehensive overview on existing scientific and non-scientific implementations of peer-to-peer information retrieval systems. In the context of this thesis, the proposed concept can be seen as a peer-to-peer system with internal and external documents (cf. Part II) that considers individual social contexts.

2.6 PRIVACY

In the information retrieval process, the act of sharing an information need and the act of replying to it unveils information about the respective senders of the information and the information need. As our later findings suggest, defining the audience for both communication forms is a critical acceptance factor (cf. Chapter 11, Chapter 16). This chapter introduces techniques to collect and operate on data in a privacy-preserving way.

2.6.1 *Privacy-Preserving Data Collection*

One of the use cases for the proposed social information retrieval system includes the collection of data from various people within one's social network to generate statistics, e.g. based on the popularity of the item (cf. Chapter 9). To avoid that the information seeker can systematically reconstruct parts of other users' information spaces by probing (Ipeirotis and Gravano, 2002) or that a user can reconcile the mapping between information item and respective information provider, it is important to offer the possibility to reply anonymously.

Simple solutions include a trusted third party that collects the information from the replying users and forwards the results to the information seeker, and at the same time removing the links to the original sender of each information item. Combining this approach with asymmetric encryption mechanisms like RSA (Rivest et al., 1978) prevents the trusted third party to read the information. In (Fung et al., 2010), the authors distinguished between record owners (the entities that have the desired information), data publishers (the entities that collect the information from the record owners), and data recipients (the entities that consume the information). In less complex cases, the data publisher is considered as a trustworthy third party, enforcing anonymity on the forwarded, aggregated information. The more complex and suitable scenario for the presented use case involves an untrusted data publisher.

Yang et al. presented an approach where a miner can collect data from a large group of people without being able to link the data to the individual respondent

(Yang et al., 2005). The protocol foresees that the group is divided into subgroups, and each subgroup sends a randomized permutation of the data to the miner, using ElGamal encryption (ElGamal, 1985) and a joint decryption technique. The approach only ensures the desired result if the message itself does not contain any information that might disclose the original author.

2.6.2 *Privacy-Preserving Set Operations*

While the information provider's need for privacy can be addressed using the techniques mentioned in the previous section, this section covers the information seeker's privacy requirements. As later studies show (cf. Chapter 11), information seekers are aware of the fact that issuing a query to their social network discloses information about themselves. Therefore, they have an interest to reduce the group of potential information providers who receive the query – apart from disguising the author of the query (e.g., using Chaum's cascade of mixes (Chaum, 1981)), a valid option would be to only ask those contacts who already have information items that would be relevant for the query. This approach would theoretically ensure that only those people get to know the query who already have information about the content domain covered by the query in their own information space and thus would not pose such a privacy risk. Therefore, it would not be possible to proclaim the query's content if one does not have content of the same type in one's individual information space. However, the two scenarios "keep information need as private as possible" and "send information need only to people who have relevant information" are not exactly the same. In a scenario where an information seeker has a privacy-sensitive information need like "tinea pedis", it might be less infringing to reveal this information need to people who have the same problem (following the logic of self-help groups: if one is there, in general one can not blame the others for being there). For the sake of the information providers' privacy needs, publishing an in-depth description of the information spaces is not considered a valid option. A solution that would fulfill both parties' privacy needs is a privacy-preserving set intersection: The information seeker has a query, represented as a feature vector $\vec{q}$. The information providers have a representation of their individual information spaces (indexed using a feature space that can be transformed to or is the same as the information seeker's feature space). Treating each dimension of the joint feature space as an item in a list if it is positive in the respective vector representation would allow to use privacy-preserving intersection protocols as proposed in (Kissner and Song, 2005). Their technique relies on a homomorphic cryptosystem and polynomials to represent multisets (multiset $S = \{S_j\}_{1 \leqslant j \leqslant k}$ is represented as polynomial $f(x) = \prod_{1 \leqslant j \leqslant k} = (x - S_j)$) and operations on multisets.

A brief example illustrates the process: assume that an information seeker A wants to send a query $\vec{q}$ to a potential information provider B. The query vector $\vec{q}$ is defined in a n-dimensional vector space, where each dimension relates to an abstract topic description. A builds a list L_A of those topic identifiers that represent $\vec{q}$ following a defined algorithm (e.g., the name of each dimension that has a positive value in $\vec{q}$). B also has a representation of her information space as a topic vector $\vec{i}$. B converts $\vec{i}$ also to a list L_B, following a similar process A already executed. Both lists are

not yet shared, i.e., A only has access to L_A and B can only see L_B. In the next step, both users follow the privacy-preserving protocol proposed in (Kissner and Song, 2005) to create the intersection of the lists (without disclosing their individual lists to the respective other party). If the intersection suggests that B might have relevant information for A, A can send a revised version of $\vec{q}$ named $\vec{q}'$ to B, where all those topic dimensions are removed which are not covered in B's information space. Following this approach ensures that B does not receive any part in A's query that is not part of B's information space and therefore constitutes a potential privacy risk for A. As already mentioned above, the described protocol is not perfect for several reasons:

- Maximizing the privacy of an information need (and therefore avoiding unnecessary sharing) is not the same as sharing the information need only with people who have relevant information. While it reduces the risk of consequences (e.g., being susceptible to blackmail) caused by the disclosure, it is no guarantee.
- It is possible to probe the information spaces of others by sending faked queries with content that could socially hurt other people. The system would identify people who have such information in their private information spaces (and therefore would be vulnerable). If probing is done on a large scale, it is easy to refuse the seriousness of the own request.
- Being able to reject the seriousness of the own request or stating that it was done with a different sentiment in mind could have the same effect as probing (and even foster such activities), rendering the protocol useless.

3

SOCIO-PSYCHOLOGICAL AND MARKET ASPECTS OF INFORMATION SHARING

Exchanging information in a social context is far more than the technical processes highlighted in the previous chapter. Human information exchange or the gathering of information has been modeled using different analogies taken from economics (cf. Section 3.1) or biology (cf. Section 3.2) to explain patterns of human behavior. In the following, a short overview of relevant work in this area is given.

3.1 INFORMATION MARKETS

Presuming that all individuals act rationally to maximize their payoffs, sharing information in a distributed social information retrieval system can also be considered as an exchange of goods in an abstract information market. The market's currency is not necessarily restricted to money and can also include information, social support or appreciation, or the endorsement to accomplish a higher-level goal. Groh and Birnkammerer proposed a market model for information in (Groh and Birnkammerer, 2011; Birnkammerer, 2010). The basic concept of their model includes transactions $T_i^{A\to C}$, where an information item i is transferred from an information broker A to a consumer C. A can be the creator of i, but this is not a formal requirement. C pays the price $P(T_i^{A\to C})$ to consume i, whereas A receives the price and the reward $R(T_i^{A\to C})$. While the price $P(T_i^{A\to C})$ reflects the explicit compensation that C gives to A, the reward $R(T_i^{A\to C})$ contains all indirect benefits that A strives to gain when C consumes i (e.g., social status, trust, awareness, etc.). Based on these ideas, Groh and Birnkammerer also derived a logic for variable and fixed costs: the costs for an information transaction can get expressed as fixed costs $K_f(A, i)$ of user A to generate or to buy i, and as variable costs $k_v(T_i^{A\to C})$, which contain variable costs for the transaction ($k_v^T(T_i^{A\to C})$, e.g., the time or resources needed to transfer or store the message) and variable costs caused by the loss of privacy ($k_v^P(T_i^{A\to C})$) that accompanies the transaction.

Azzopardi et al. proposed an economic model for information retrieval (Azzopardi, 2011; Azzopardi et al., 2013; Azzopardi, 2014). Unlike Groh and Birnkammerer's model, Azzopardi et al. focus on conventional search engines and analyze the information seeker's cost to process the results and to pose queries. Azzopardi et al.'s model therefore does only cover the side of the consumer in the information market. One of Azzopardi et al.'s findings is that with increasing costs for queries, users tend to issue fewer queries and examine more documents per query (Azzopardi et al., 2013).

Other research published in (Bertino and Matei, 2015) covers the perspective of the information provider and thus provides insights into reputation, trust, credibility, and different forms of contribution in social media scenarios.

Adler and Kwon (Adler and Kwon, 2002) provide a comprehensive overview of definitions for "social capital" from a sociological perspective and distinguish between market relations, hierarchical relations, and social relations. As already mentioned in Section 2.4.2, Fiske proposed a framework to explain social relations with four psychological models (Fiske, 1992): communal sharing (CS), authority ranking (AR), equality matching (EM), and market pricing (MP). Communal sharing describes a social interaction mode where all members of a group treat each other as equivalent. Authority ranking linearly orders people according to their (social) positions. In equality matching, "people keep track of the imbalances among them" (Fiske, 1992) and in market pricing, interactions are reduced to a one-dimensional value.

3.2 INFORMATION FORAGING TECHNIQUES

Pirolli and Card (Pirolli and Card, 1995; Pirolli, 2009) presented an information foraging model that tries to explain human's behavior when satisfying information needs. The theory behind the model "derives from optimal foraging theory in biology and anthropology, which analyzes the adaptive value of food-foraging strategies. Information foraging theory analyzes trade-offs in the value of information gained against the costs of performing activity in human-computer interaction tasks" (Pirolli and Card, 1995). The value of the information is not solely related to the information itself, but can only be assessed in the context of the embedding task. Pirolli stated that low-cost information foraging behaviors (i.e., taking the obvious results) are associated with the core zones of the content domain, while high-cost information foraging behaviors are associated with peripheral zones (i.e., require more effort) (Pirolli, 2009). Furthermore, Pirolli supported Granovetter's idea of the positive effects of weak ties (cf. Section 2.5.1) and concluded that "[homogeneity] of opinion, viewpoint, and information resources among a group of information foragers is likely to produce redundancy in what they find and how they interpret those findings. We might expect that groups of cooperative information foragers will be more effective if constituted by individuals with some degree of diversity" (Pirolli, 2009). Following Pirolli's theory, information seekers evaluate the utility of the information that has already been processed and continue to search if the expected additional gain is higher than the costs of the search process.

4

FUNDAMENTALS OF SEMANTIC WEB

In Chapter 2, information retrieval was introduced. While information retrieval deals with finding information about a subject, *data retrieval* "aims at retrieving all objects which satisfy clearly defined conditions such as those in a regular expression or in a relational algebra expression" (Baeza-Yates and Ribeiro-Neto, 1999, p. 1). Information retrieval techniques typically operate on unstructured data (like natural-language document collections without an explicit semantic structure), whereas data retrieval approaches need a well-defined structure in the data. Data retrieval systems operate on higher-order predicate logic and can be implemented using database systems. In information retrieval, relevance is a continuous measure, while it is binary (or ordinal) in a data retrieval scenario. For a data retrieval system, "a single erroneous object among a thousand retrieved objects means total failure[,] (...) [while] for an information retrieval system, (...) the retrieved objects might be inaccurate and small errors are likely to go unnoticed" (Baeza-Yates and Ribeiro-Neto, 1999, p. 2). Semantic web is interpreted as a technology bundle to explicitly represent and semantically annotate data. Thereby, semantic web technologies allow to apply data retrieval approaches. One of the examples implemented in the present thesis (Social Product Search, cf. Section 16.6.3) operates in the data retrieval domain and therefore partially relies on semantic web concepts.

According to (World Wide Web Consortium, 2013), the "Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. (...) It is based on the Resource Description Framework (RDF)". The term was coined in 2001 by Berners-Lee et al. (Berners-Lee et al., 2001) and describes the idea that data is linked to a semantic interpretation that is understood by computers. It therefore builds upon the achievements of previous research in the (symbolic) Artificial Intelligence (AI) domain (Shadbolt et al., 2006) (and forms a counterpart to Machine Learning (Murphy, 2012; Bishop, 2006), where explicit symbols are omitted). The basic idea is that data is annotated with additional meta-information, which is explicitly defined in metadata models like RDF[1], OWL[2], or XML[3].

Those technologies allow to explicitly model relationships between items and therefore – in theory – conduct mainly conceptual reasoning. One requirement to efficiently use semantic web technologies is the existence of explicit semantic annotations (linked to a taxonomy or ontology).

In the domain of products, which is relevant in some of the use cases of the concept developed in this thesis, several possible taxonomies exist. The most established system to categorize products is the one maintained by GS1[4]. GS1 is a non-profit organization and controls the allocation of Global Trade Item Numbers (GTIN) as

1 http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/ (retrieved 2015-11-20)
2 http://www.w3.org/TR/owl-ref/ (retrieved 2015-11-20)
3 http://www.w3.org/TR/2006/REC-xml11-20060816/ (retrieved 2015-11-20)
4 http://www.gs1.org/ (retrieved 2015-11-20)

a unique identifier for products. GS1 works cross-industry and is the dominant solution for retailers[5]. GS1 provides a comprehensive category schema for products called Global Product Classification (GPC)[6]. A free service to query this database is available at `http://opengtindb.org/`[7].

As a pragmatic low-level approach, AMAZON'S ASIN SYSTEM can be used[8]. Each ASIN consists of ten letters and/or numbers and can be seen as a unique identifier for a product listed in Amazon's catalogue (ASIN is the abbreviation for Amazon Standard Identification Number). Using the Product Advertising API of Amazon Web Services[9], it is possible to retrieve additional meta-information for each product, including the respective category in Amazon's product category system. Furthermore, it is possible to retrieve the Amazon website for an ASIN using the URL *http://www.amazon.com/dp/<ASIN>*. Amazon's category system consists of several layers, e.g. *Hasbro's Monopoly* with ASIN *Booo05N5PF* is located in *Toys & Games* → *Games* → *Board Games*[10].

Semantic Web Technologies have already been used in social contexts for Social Semantic Desktop Environments (Groza et al., 2007). The approach tries to define a common metadata format to allow a seamless integration of different tools and workspaces, including collaboration and social interaction. The focus of the present thesis is different: we try to identify benefits of integrating the social environment into the search process. Therefore, a concept which allows modeling a wide range of social information retrieval approaches is proposed and prototypically implemented. The investigated scenarios range from a purely manual approach (offering the highest flexibility due to the manual interaction with the information provider, cf. Section 16.5.2) to an automated approach, where a digital information space is indexed and searched (cf. Section 16.5.3). A special use case includes a product search scenario, that also relies on data retrieval techniques (cf. Section 16.5.4). For all scenarios, the objective is not to prove technical feasibility or optimal performance measures, but to distill (technical and social) insights from the different modes which might be useful for future implementations of social information or data retrieval systems.

5 `https://en.wikipedia.org/wiki/Global_Trade_Item_Number` (retrieved 2015-11-20)
6 `http://www.gs1.org/how-gpc-works` (retrieved 2015-11-20)
7 (retrieved 2015-11-20)
8 `http://www.amazon.com/gp/seller/asin-upc-isbn-info.html` (retrieved 2015-11-20)
9 `http://docs.aws.amazon.com/AWSECommerceService/latest/GSG/Welcome.html`, `http://aws.amazon.com/` (retrieved 2015-11-20)
10 `http://www.amazon.com/dp/B00005N5PF` (retrieved 2015-11-20)

5

STATISTICS & TOOLS

In the following sections, the statistical tools and methods used for the experiments' evaluation are introduced briefly. For a more in-depth description of the topics, please refer to the cited literature.

5.1 CORRELATION COEFFICIENTS

5.1.1 *Pearson's Correlation Coefficient*

Pearson's correlation coefficient (also referred to as Pearson's r) is defined for two sample datasets $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_n\}$ as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{7}$$

where $\bar{x} = \frac{1}{n}\sum_i x_i$ and $\bar{y} = \frac{1}{n}\sum_i y_i$. The result r can be interpreted as a measure to quantify the degree and the direction of a linear correlation between the two samples. It ranges from -1 (strong negative linear correlation) to 1 (strong positive linear relation). If $r = 0$, both datasets do not correlate linearly (Runkler, 2015).

Pearson's r requires x and y to be on an interval scale and approximately normally distributed. Furthermore, x and y must be linearly correlated. The significance level of r can be measured using a t-test with the following test metric

$$t = \frac{r}{1 - r^2} \cdot \sqrt{n-2} = r \cdot \sqrt{\frac{n-2}{1-r^2}} \tag{8}$$

with r denoting Pearson's r and n being the number of pairs in both datasets. If both samples follow a nearly normal distribution, t follows a t-distribution with $(n-2)$ degrees of freedom. If the absolute value of t is greater than the respective value of the t-distribution in a two-tailed test for a given probability t and $(n-2)$ degrees of freedom, the result is statistically significant at this level.

5.1.2 *Spearman's Correlation Coefficient*

Spearman's correlation coefficient (also referred to as Spearman's rho) is a non-linear measure of correlation between two samples $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_n\}$. A completely positive (rho=1) or negative (rho=-1) correlation exists when each variable is a perfect monotonic function of the other. The n values of each of the two datasets ($\{x_1, ..., x_n\}$, $\{y_1, ..., y_n\}$) are converted to ranks ($\{r_{x,1}, ..., r_{x,n}\}$, $\{r_{y,1}, ..., r_{y,n}\}$) and rho is calculated as follows:

$$\text{rho} = 1 - \frac{6\sum_i (r_{x,i} - r_{y,i})^2}{n(n^2-1)}. \tag{9}$$

The significance of rho can be calculated using

$$t = \text{rho} \cdot \sqrt{\frac{n-2}{1-r^2}} \tag{10}$$

which also follows a t-distribution with $n-2$ degrees of freedom. Compared to Pearson's r, Spearman's rho is less sensitive to outliers, because all values are converted to ranks before (and thereby limiting outliers to the value of their ranks, while Pearson's r considers the full value and is therefore more prone to outliers) (Myers et al., 2010, p. 485).

5.2 REGRESSION

5.2.1 *Linear Regression*

CONCEPT Given a continuous response variable (also referred to as dependent variable) y and a vector of predictor variables (also referred to as independent variables or explanatory variables) $\vec{x}$, linear regression aims to express the relationship using a linear function

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_{k+1} x_{k+1} + \epsilon \tag{11}$$

β_0 denotes the intercept and $\beta_{i>0}$ is the estimated coefficient of the explanatory variable x_i. ϵ quantifies the prediction error and is called the residual (and is individual for each measurement and normally distributed). The model parameters β_i (with $i \in \{0, ..., k+1\}$ are estimated by minimizing the sum of squared errors for all measured and predicted values of y (i.e., $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ with y_i being the actual and $\hat{y}_i$ being the predicted value of the i-th measurement). According to (Gelman and Hill, 2007), the least squares estimate equals the maximum likelihood estimate "if the errors ϵ are independent with equal variance and normally distributed" (Gelman and Hill, 2007, p. 40).

In Machine Learning, the concept of "traditional" linear regression is extended by using different basis functions (Bishop, 2006, p. 138). The more general version of linear regression can therefore be defined as

$$y(x) = \beta_0 + \sum_{i=1}^{k+1} \beta_i \phi_i(x) + \epsilon \tag{12}$$

with $k+1$ being the total number of parameters in the model and $\phi_i(x)$ denoting the basis functions. The basis functions can be nonlinear, allowing $y(x)$ "to be a nonlinear function of the input vector x". The model is still considered to be a linear model, as it is linear in β_i (Bishop, 2006, p. 139). In standard linear regression (as shown in Equation 11 above), the general basis function $\phi_i(x)$ has been defined as $\phi_i(x) = x^i$. Furthermore, machine learning literature commonly refers to the coefficients as $\mathbf{w}$ instead of β (Murphy, 2012, p. 19).

REQUIREMENTS Linear regression is based on a set of assumptions with regard to the data, which are discussed controversial. (Osborne and Waters, 2002) and (Gelman and Hill, 2007, p. 45) provide a very comprehensive and recent agreement. Following Gelman and Hill's argumentation, a regression model should meet the following criteria (in decreasing order of importance):

- *Validity*: The model data should fit to the research question, i.e. the response variable should map to the phenomenon of interest and the explanatory variables should cover all relevant predictors.
- *Additivity and linearity*: The response variable should be a linear function of the explanatory variables. If additivity is violated, transforming the variables might be useful (e.g., $y = a \cdot b \cdot c$ can be transformed to $\log(y) = \log(a) + \log(b) + \log(c)$). Other commonly used functions to create linear dependencies or stabilize variance include exp, $1/y$ or $\sqrt{y}$ (Crawley, 2007, p. 205).
- *Independence of errors*: Residuals should be independent from each other.
- *Equal variance of errors*: The variance of the residuals should be constant (i.e., homoscedastic). If the variance of the errors is heteroscedastic, "estimation is more efficiently performed using weighted least squares, where each point is weighted inversely proportional to its variance (...). In most cases, however, this issue is minor. Unequal variance does not affect the most important aspect of a regression model" (Gelman and Hill, 2007, p. 45).
- *Normality of errors*: Unlike many older textbooks, Gelman and Hill do not recommend to check the regression residuals for normality (Gelman and Hill, 2007, p. 45).

INTERPRETATION Conventional statistical software tools like R[1] provide additional information for linear models: for the intercept β_0 and each coefficient β_j, an estimate is calculated along with the expected standard error. The estimated coefficient divided by the standard error leads to the t-value, which is then used to calculate the p-value based on the t-distribution. The p-value can be interpreted as probability that this t-value or a larger one can get obtained by chance alone (Schäfer, 2011; Crawley, 2007).

EVALUATION A metric often used to assess the quality of a linear model with multiple explanatory variables is R^2 (often also referred to as multiple R^2). It is defined as $\frac{\text{Var}(\hat{y})}{\text{Var}(y)}$, with Var(.) being the variance, $\hat{y}$ representing the fitted values of the regression model, and y referring to the actual values of the response variable. R^2 can be interpreted as the amount of variance that is explained by the regression model. If the regression model only contains a single variable, it is written in lowercase (r^2). A more reliable measure to evaluate the quality of a model is adjusted R^2, defined as

$$\text{adj. } R^2 = R^2 - (1 - R^2)\frac{p}{n - p - 1} \tag{13}$$

1 The R Project for Statistical Computing, https://www.r-project.org/ (retrieved 2015-11-23)

with p denoting the number of independent variables.

To optimize models, the *Akaike Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC) are commonly used metrics to compare two different models (cf. (Burnham and Anderson, 2004) for an in-depth comparison).

5.2.2 *Logistic Regression*

CONCEPT Logistic regression models are based on the logistic function $\frac{\exp(x)}{1+\exp(x)}$ and predict a binary response variable. Therefore, the error distribution is binomial (Crawley, 2007, p. 514). For a binary response variable $Y_i \in \{0, 1\}$ and a k-dimensional explanatory variable X the model is defined as (cf. (Wasserman, 2005, p. 223))

$$p_i = P(Y_i = 1|X = x) = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j})}, \tag{14}$$

and can get expressed as

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} \tag{15}$$

with

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right). \tag{16}$$

A basic concept in logistic regression are odds ratios. The odds for a certain event $Y_i = 1$ are defined as

$$\frac{P(Y_i = 1|X = x)}{1 - P(Y_i = 1|X = x)} = \frac{p_i}{1 - p_i}. \tag{17}$$

To have a linear relationship in the model, the model is transformed using $\log(.)$, leading to *logits* (as defined above). Logistic regression models estimate regression weights for each explanatory variable to calculate the respective logits.

INTERPRETATION The direct interpretation of the estimated coefficients β_i in Equation 15 is difficult. Therefore, the values are first transformed back to normal odds ratios: as in simple linear regression, β_0 is the intercept and can be interpreted as the predicted log of the odds for p_i if all explanatory variables are 0 (i.e., if $\forall j : x_j = 0$, $\frac{p_i}{1-p_i} = \exp(\beta_0)$). A similar logic applies to the coefficients: keeping all other explanatory variables constant, a one-step increase of explanatory variable $x_{i,j}$ would lead to an estimated change of the odds for p_i to $\frac{p_i}{(1-p_i)} \cdot \exp(\beta_j)$. A more in-depth introduction to logistic regression is given in (Wasserman, 2005, p. 223), (Gelman and Hill, 2007, p. 79), or (Crawley, 2007, p. 569). In the machine learning context, logistic regression is interpreted as a (binary) classifier (Murphy, 2012, p. 21).

EVALUATION The performance of a logistic regression model can get evaluated when comparing it with the nonparametric null model (the model only assigning the same probability to each y_i based on the proportion in the measured data). The metric to compare the errors in both models is *deviance* (cf. (Gelman and Hill, 2007, p. 100) for more details):

- When an explanatory variable with random noise is added to the model, deviance should increase by 1.
- When an informative explanatory variable is added to the model, deviance should decrease by more than 1.

Formally, *deviance* is defined as $-2(L(\mu, y) - L(\nu, y))$, with L denoting the log-likelihood function, y being the observed data, and μ and ν being the predicted data of two models (Agresti, 2002, p. 118). One of the models often refers to the saturated model, i.e. the model that fits exactly the observed data, because it contains a parameter for each measurement. Another commonly used approach is to compare a model to the null model (see above).

As for the simple linear regression model, several R^2 measures are defined for logistic regression models to calculate a "goodness of fit" value. Long listed the commonly used McFadden pseudo-R^2 metric and its adjusted version which compensates for models with many explanatory variables (Long, 1997, p. 104).

5.2.3 *Ordinal Logistic Regression*

CONCEPT Ordinal logistic regression can be interpreted as a generalized form of logistic regression, where the response variable is on an ordinal scale. In this thesis, cumulative link models (also referred to as proportional odds models) are used. An introduction for the specific implementation can be found in (Christensen, 2015b; Christensen, 2015a). As already defined above, the logistic regression model can be written as

$$\log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{P(Y_i = 1|X)}{1 - P(Y_i = 1|X)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} \tag{18}$$

and is based on the odds that the event $P(Y_i = 1|X)$ occurs (with $Y_i \in \{0, 1\}$ being a binary variable). For an ordinal response variable, the event of interest is modeled to reach a specific score on the ordinal scale or less. Imagine an ordinal response variable Y which consists of four levels: (1) poor, (2) fair, (3) good, and (4) excellent. The odds of the events follow the form $\theta_j = P(Y_i = j|X)/P(Y_i > j|X)$ and are defined as

- $\theta_1 = P(Y_i = 1|X)/P(Y_i > 1|X)$,
- $\theta_2 = P(Y_i \leqslant 2|X)/P(Y_i > 2|X)$, and
- $\theta_3 = P(Y_i \leqslant 3|X)/P(Y_i > 3|X)$.

θ_4 does not exist in the example introduced above because $P(Y_i \leqslant 4|X) = 1$ and $P(Y_i > 4|X) = 0$. For each independent variable, the ordinal logistic regression model is

$$\log\left(\theta_j\right) = \alpha_j - \beta X \tag{19}$$

with j ranging from 1 to the number of categories minus 1 (Norušis, 2011, p. 71). To calculate the coefficients β, a maximum likelihood estimation is used in the cumulative link model (with logit link) presented in (Christensen, 2015b). The model can easily get extended to multiple explanatory variables: with an ordinal response variable Y with c category levels and a k-dimensional explanatory variable vector X, the cumulative link model (also referred to as proportional odds model, cumulative logit model, and ordered logit model) is defined as

$$\operatorname{logit} P\left(Y_i \leqslant j|X_i\right) = \alpha_j - \sum_{l=1}^{k} \beta_l x_{i,l} \tag{20}$$

for category $j < (c-1)$ with $i \in \{0, n\}$ being the index for the n data pairs $(y_i, \vec{x_i})$ and $x_{i,l}$ referring to the l-th component of the vector $\vec{x}_i$. α_j is the intercept for each category level of Y, while β_j is the respective intercept of that category level (cf. (Agresti, 2002, p. 275) and (Christensen, 2015b)).

INTERPRETATION As in the logistic regression model, the coefficients β_l can get interpreted in the following way: when keeping all other factors constant, a single step increase of explanatory variable $x_{i,l}$ changes to odds to reach the next higher category of Y by a multiplicative factor of $\exp(\beta_l)$ (UCLA: Statistical Consulting Group, 2014).

EVALUATION As in the logistic regression model, the performance of an ordinal logistic regression model is evaluated by comparing it to the null model, i.e. the nonparametric model using deviance.

5.2.4 *Random Effects Models*

Random effects models are a way of multilevel modeling and explain a predicted variable with *fixed effects* (using independent variables as predictors, as mentioned above in the other models) and additional random effects caused by uncorrelated quantities (which were obtained e.g. by multiple measurements due to individual preferences of the regarded subject). Random effect models resemble the previously mentioned models above, with one significant difference: in addition to the observed heterogeneity explained by the independent variables in the fixed effects models, the unexplained heterogeneity is considered as a random effect caused by quantities orthogonal to the independent variables. Random effects models account for the variances between groups in the intercept or the slope of the model (varying-intercept and varying-slope model). As stated above, a trivial standard linear regression model can be described as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{21}$$

Assuming that the model is fitted on data which can be split in groups J, a varying-intercept model can be written as

$$y_i = \beta_{0,j} + \beta_1 x_i + \epsilon_i \tag{22}$$

(with $j \in J$), while a varying-slope model reflects the groups in β_1

$$y_i = \beta_0 + \beta_{1,j} x_i + \epsilon_i. \tag{23}$$

A random effects model can therefore be seen as a group of fixed effect models, taking into account particularities of the groups in the measured data. The model is referred to as "random effects model" because the group-specific parameters (i.e., $\beta_{0,j}$, $\beta_{1,j}$ in the examples above), are estimated using a normal distribution fitted to the observations in the data. For a more detailed introduction, please refer to (Gelman and Hill, 2007, p. 237) or (Crawley, 2007, p. 627). In the thesis, random effect models (with varying intercept) are used to account for repeated measurements.

5.2.5 *LOESS regression*

LOESS regression is a technique to fit a regression line based on local smoothing (Cleveland and Devlin, 1988). In contrast to regression approaches based on a global function for the whole model, LOESS allows to fit the data locally, i.e. use the best approximation for each subset of data without being forced to align it with a global function. Drawbacks of this approach include the increased computation effort and the inability to represent the result in a closed form, e.g. a function. LOESS regression is useful for data exploration to detect relations between variables without being forced to commit to a specific function being fitted to the measurements.

5.3 STATISTICAL TESTS AND METHODS

5.3.1 *Analysis of Variance (ANOVA)*

ONE-WAY LAYOUT In the one-way layout, the analyzed data consists of a measured value (Y) and a (single) factor that splits the samples into several groups. For the analysis, three different variances are of interest:

- Variance across all samples (this is the variance one tries to explain),
- variance within each group (groups are defined by the factor), and
- variance between the groups.

To assess how much variance is explained by the grouping factor, the variance between the groups (defined by the factor, i.e. the explained variance) has to be related to the variance within the group (variance not explained by the factor, i.e. caused by other sources). For the result, it is possible to calculate a significance value using the F-distribution. For a detailed introduction please refer to (Schäfer, 2011, p. 117) and (Rice, 2007, p. 477). It is important to note that error components need to be normally distributed and that error variances between the groups must be homogeneous (homoscedasticity).

TWO-WAY LAYOUT ANOVA is also possible with more than one explanatory variable. For a detailed description, please refer to (Schäfer, 2011, p. 125).

5.3.2 *Wilcoxon Rank Sum Test*

WILCOXON RANK SUM TEST (also referred to as *Mann-Whitney U test*, *Wilcoxon Mann Whitney test*, *U-test*) is a nonparametric test to verify that two populations are different. In contrast to the t-test, it does not require a Gaussian distribution of the data. H_0 denotes that the two distributions do not differ, while H_1 denotes the hypothesis that the datasets are different. As a test statistic, it is possible to use the Mann-Whitney-U statistic or the Wilcoxon-Rank-Sum statistic W since both can be transformed into each other (Schäfer, 2011, p. 143), (Crawley, 2007, p. 297), (Rice, 2007, p. 435).

5.3.3 *Kruskal-Wallis Test*

The KRUSKAL-WALLIS TEST is a generalization of the Wilcoxon rank sum test introduced above for an ordinal variable. In contrast to the Wilcoxon rank sum test, the Kruskal-Wallis test allows to compare more than two groups (defined by the ordinal variable). As in the Wilcoxon rank sum test, H_0 is defined as "no difference between the groups" (Kruskal and Wallis, 1952),(Rice, 2007, p. 488).

5.3.4 *Durbin-Watson Test*

The DURBIN-WATSON TEST (Durbin and Watson, 1971) tests for autocorrelation of the residuals in a regression model (H_0: residuals are uncorrelated). To improve the quality of a linear regression model, autocorrelation should be avoided.

5.3.5 *Testing for Normality*

SHAPIRO-WILK NORMALITY TEST Shapiro and Wilk proposed a test for normality based on the analysis of variance (H_0 assumes that the data is normally distributed) (Shapiro and Wilk, 1965).

5.3.6 *Testing for Heteroscedasticity*

As mentioned in Section 5.2.1, one of the requirements for linear regression is the homogeneous variance of the residuals. Breusch and Pagan developed a method to test for heteroscedasticity (Breusch and Pagan, 1979). The H_0 hypothesis states that the variance of the tested data is homogeneous (i.e., the data is *homoscedastic*). The test is used to verify the homoscedasticity requirement for the residuals in linear regression and requires a certain degree of normality in the residuals.

5.4 TOPIC MODELS

There are conceptual differences between traditional statistical methods and prediction techniques from the machine learning domain. We will not further elaborate on these differences, but introduce *Topic Models* as a generative model from the machine learning domain.

Topic models identify a semantic structure in a text corpus and therefore offer a possibility to detect recurring themes or latent *topics* in huge data collections. Prominent examples are Latent Semantic Indexing (LSI) proposed by Deerwester et al. (Deerwester et al., 1990) based on Singular Vector Decomposition (SVD) to reduce the dimension of the underlying vector space (and to avoid synonymy and polysemy). Later, Hofmann introduced Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999). While LSI is based on linear algebra and reduces the occurrence matrix via SVD, pLSI uses a latent-class model to model each document as a mixture of topics. With Latent Dirichlet Allocation (LDA), Blei et al. proposed a new topic model approach based on pLSI (Blei et al., 2003). The basic generative process is very similar to the one pLSI uses, however, while pLSI's mixture of topics is conditioned on each document, LDA's topic mixture is modeled using a constant Dirichlet prior for all documents (Wei, 2007). According to (Wei, 2007), "[LDA] has quickly become one of the most popular probabilistic text modeling techniques in machine learning and has inspired a series of research papers" (Wei and Croft, 2006; Baillie et al., 2011). Therefore, it will be explained in more detail in the following section.

5.4.1 *Latent Dirichlet Allocation (LDA)*

In this section, the basic mechanisms of LDA will be explained following (Blei, 2012; Blei et al., 2003). The basic idea of LDA is that documents cover multiple topics. Each topic is defined as a distribution of words that characterizes the topic (e.g., literature covering computer science topics contains different terms than medical texts). LDA is a generative probabilistic model, so it is defined in a way that it can create the observed data. For the initial training phase, the perspective is flipped: using the documents, the parameters of the model are adjusted (using Maximum Likelihood Estimation) in such a way that it is most likely that the respective model has created the documents. Commonly used techniques include Variational Bayesian Methods (Blei et al., 2003), Gibbs Sampling (using Markov chains, (Geman and Geman, 1984)), and Expectation Propagation (Minka and Lafferty, 2002).

More formally, LDA sticks to the following process to generate each document w in a corpus D (cf. (Blei et al., 2003)):

1. Choose N (length of w) from the Poisson distribution $\mathtt{Poisson}(\xi)$
2. Choose θ (distribution of topics for the current document) from the Dirichlet distribution $\mathtt{Dir}(\alpha)$
3. For each of the N words w_n of the document:
   a) Choose topic z_n from a multinomial distribution $\mathtt{Multinomial}(\theta)$
   b) Choose a word w_n from $P(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

As a consequence, each of the N generated words is conditioned on parameter β and the topic z_n, and each topic z_n is conditional dependent on the topic mixture θ. α is the parameter of the Dirichlet prior on the distributions for each document, β is the parameter of the Dirichlet prior for the topic-word distribution, θ_i represents the topic proportions for document i, $z_{i,n}$ is the topic assignment for the n-th word in i, and $w_{i,n}$ denotes the n-th word in document i (which is an element of the fixed vocabulary and the only observed variable, the other variables are latent) (Blei et al., 2003; Blei, 2012).

With a learned topic model, it is possible to transform previously unseen documents to the topic space and get a measure how close a document is to each of the existing topics (using θ for the respective document).

5.4.2 *Similarity Measures*

Documents are represented in the topic space using a discrete probability distribution over the topics in the corpus. This distribution specifies to which degree the document is related to the respective topics. Therefore, it is possible to estimate the similarity of two documents by calculating their representations' similarity in the topic space. In this section, two prominent approaches are introduced.

HELLINGER DISTANCE Hellinger Distance is a distance measure to quantify the distance between two probability distributions. For two documents d and f, it is defined as

$$\mathrm{HD}_{d,f} = \sum_{k=1}^{K} \left(\sqrt{\hat{\theta}_{d,k}} - \sqrt{\hat{\theta}_{f,k}} \right)^2 \tag{24}$$

with K being the total number of topics and $\hat{\theta}_{d,k}$ being the posterior topic proportion for a given document d and topic k (Blei and Lafferty, 2009).

JENSEN-SHANNON DIVERGENCE Jensen-Shannon Divergence is a distance measure to compare probability distributions and is therefore suitable to compare documents represented by discrete probability distributions over topics. For two documents d and f, it is defined as

$$JSD_{d,f} = \frac{1}{2}KLD\left(\hat{\theta}_{d,k}, M_{d,f}\right) + \frac{1}{2}KLD\left(\hat{\theta}_{f,k}, M_{d,f}\right) \tag{25}$$

with

$$M_{d,f} = \frac{1}{2}\left(\hat{\theta}_{d,k} + \hat{\theta}_{f,k}\right) \tag{26}$$

and

$$KLD(d,f) = \sum_{k=1}^{K} \hat{\theta}_{d,k} log \frac{\hat{\theta}_{d,k}}{\hat{\theta}_{f,k}} \tag{27}$$

being the Kullback-Leibler Divergence, K referring to the total number of topics, and $\hat{\theta}_{d,k}$ denoting the posterior topic proportion of a document d and a topic k. *Jensen-Shannon distance* is a metric defined as the square root of the Jensen-Shannon divergence (Endres and Schindelin, 2003).

5.5 EXPLICIT SEMANTIC ANALYSIS (ESA)

Explicit Semantic Analysis (ESA) is an approach to represent text in a vector space defined by Wikipedia concepts. Given a vector of basic concepts (the Wikipedia article pages), $C_1, ..., C_m$, each term t can be represented as a vector of weights $w_1, ..., w_m$, with each weight w_i expressing the degree of relatedness of the term t to the respective concept C_i. To calculate the degree of relatedness between t and the concepts, TF-IDF is used:

$$w_i = tf(t, C_i) \cdot log \frac{m}{df_i} \tag{28}$$

df_i denotes the number of concepts that contain the term t and as above, m is the total number of concepts. The concept C_i refers to the article describing C_i. The vector of weights $\vec{w}$ could be seen as an interpretation vector. For a term "mouse", it would have two strong components, which correspond to the two possible meanings "mouse (rodent)" and "mouse (computing)". The same would apply to the term "screen", being interpreted as "computer screen" and "window screen". When interpreting a sentence like "I purchased a mouse and a screen", the respective computing components would add up and boost the interpretation towards computer-related components, which would effectively support disambiguation (Gabrilovich and Markovitch, 2009). Gabrilovich and Markovitch also describe a sliding-window approach to reduce the noise in the data by setting those relations between terms and concepts to zero where the relation is too spurious. For details, please refer to (Gabrilovich and Markovitch, 2009, p. 453).

Part II

CONCEPT FOR A SOCIAL INFORMATION RETRIEVAL SYSTEM

In the following chapters, an exemplary concept for a social search system is presented. The main objective of the concept is to evaluate limits and chances of Social Information Retrieval. It therefore covers different specific scenarios for interaction patterns, data organization, or routing to allow conclusions for a broad range of Social Information Retrieval approaches. The proposed concept is highly variable and can be parameterized, while being sufficiently specific to define our understanding of Social Information Retrieval. It also represents a confinement of the thesis' object of study and provides a valid solution to explore the effects of interest. While the concept is build on our understanding of the state-of-the-art in the problem domain, it does not necessarily claim to be the optimal solution – far more important is that the approach is elaborated enough to allow a profound investigation of the research questions. Chapter 6 explains the architecture and basic functionality of the system. Chapter 7 elaborates on the routing mechanisms used to identify suitable information providers based on the information seeker's social network, the advertised knowledge of the information providers, and previous interactions. In addition, an example of a privacy-aware social interaction protocol is introduced as a proof of concept, which allows the information seeker and the information provider to ask or reply to questions without unveiling their identity. Chapter 8 describes various ways to organize information spaces and to extract information from the original systems. Chapter 9 closes this part of the thesis with a description of four exemplary use cases where a social information retrieval system would be beneficial for the users.

6

ARCHITECTURE

The fundamental idea of the proposed social information retrieval concept allows users to leverage information that tacitly exists in their social environment. Users can send queries to other people in their social network. Supported by the system, the recipient of such a query can reply with an information item taken from her private information space and/or with a manually written reply to the query. The concept resembles traditional offline information retrieval conducted by humans: by talking to each other, people pass information, using an individual authorization schema defining which information should be shared with whom. By relying on the information seeker's social network, contextual factors could increase the chance to retrieve results with a higher "social relevance" than traditional search engines could offer, i.e. items that are relevant because they are relevant within one's social network. This could foster the discovery of new items that would not have been considered as relevant before (serendipity, Section 2.3.5). For a significant share of queries this "social relevance" is not of primary interest (e.g., for navigational queries), but related work (Lee and Brusilovsky, 2012; Groh et al., 2011a; Groh and Ehmig, 2007) in the recommender systems domain suggests that specific types of information needs could benefit from this concept. In addition, social relationships could increase the available content for information retrieval and therefore represent a way to access the hidden web (Kautz et al., 1997b): users might not mind to share unpublished private information with socially close information seekers if they know that they can help to satisfy their information need. On the other hand, focusing only on the content that is available in the information spaces of socially close peers, limits the available information. For certain information needs, it could be valid to assume that the additional gain from the peers' information spaces weights higher than the loss from restricting oneself to the content available in one's own social network (especially when considering social information retrieval as additional method to satisfy information needs).

Each participant of the system maintains her own private information space. This information space consists of (but is not necessarily limited to) the documents a user read, the user's written communication, and the transactional data a user generated. Connectors could e.g. add email messages from the user's email account or the content of browsed websites to the indexed information space. Each action conducted by the user extends the user's indexed information space, provided that a connector is able to create an item in the indexed information space reflecting the action and/or its outcome. Examples for actions are writing reports, buying products, or searching and browsing the web. Chapter 8 explains the concept of information spaces in greater detail, Section 8.1 describes the representation of knowledge in the indexed information space, and Section 8.2 elaborates on the extraction process of information from the original systems.

The system's central component is a user agent that can get controlled using a frontend accessible from the user's devices (e.g., cell phone, tablet computer, laptop,

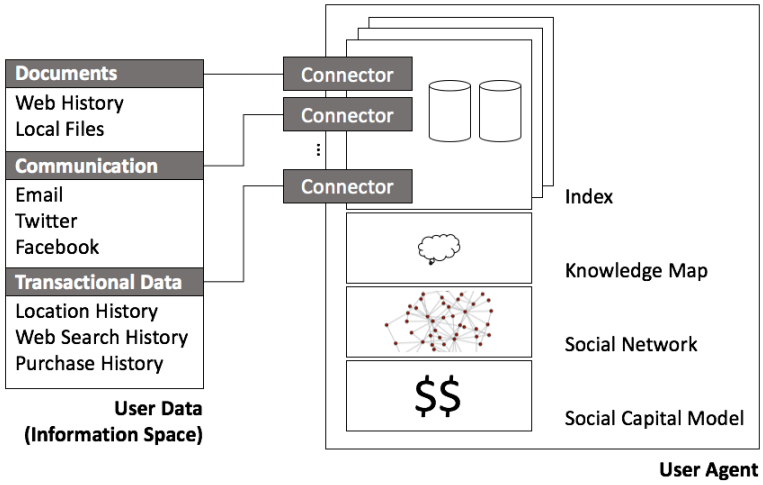


Figure 4: Schematic structure of the user agent

stationary computer, etc.). The data stored on the agent is kept synchronized among the devices of the user, e.g. by storing a copy on the internet using well-established synchronization protocols.

Apart from maintaining the indexed information space, the user agent also keeps a map of other users who might be knowledgeable in a certain topic area, a representation of the user's social network, and a model for the user's social capital transactions. Figure 4 shows the user agent's schematic overview.

Each user can query the information spaces of other users by sending queries to them (phrased as a human-readable question and/or traditional search terms, depending on the mode). The user who has issued the query is referred to as "Information Seeker" (IS), a user who has been selected as recipient for the query is referred to as "Designated Information Provider" (DIP), while a user who actually replied to the query is denoted as "Information Provider" (IP). In Chapter 7, different query modes are introduced (Section 7.2) and the process to find an appropriate set DIPs is explained (Section 7.1).

When the agent of a DIP receives a new query, several potentially relevant information items from the DIP's private information space are assembled. This list is presented to the DIP as a recommendation for a reply to the query. The DIP can adjust the reply (e.g., add/remove information items or enter a different response manually) and send it back to the IS or decide to ignore the request. Based on the configuration of the information space, the relationship between IS and DIP, and the type and privacy attributes of the respective information items, the identified information items could also get automatically shared with the IS.

In the following, Chapter 7 explains the definition of queries (Section 7.2), a way to identify a suitable set of recipients for a query (Section 7.1), and a social capital model to economically model information exchange (Section 7.1.3). Chapter 8 details the concept of the information space, including the representation (Section 8.1) and extraction (Section 8.2) of information. Chapter 9 closes this part of the thesis with an overview of possible use case scenarios.

7

QUERY DEFINITION AND ROUTING

7.1 FINDING THE RIGHT INFORMATION PROVIDER

Three components of the system define the proposed set of designated information providers for an information seeker's query:

- The *Social Network* component covers several attributes of the social edges between two users, i.e. tie strength, sympathy, similarity of social context, or similarity of content knowledge.
- The *Knowledge* component represents the content categories where other users are considered knowledgeable (as far as the information seeker is aware).
- The *Social Capital* component quantifies previous interactions using an economic model.

Each component will give a vote how well a potential candidate would fit to the query in the respective dimension (social network, knowledge, social capital). Inspired by (Macdonald and Ounis, 2006), a linear combination of all votes, multiplied with empirically evaluated weighting factors for each dimension, will lead to a one-dimensional result for each potential information provider. With $SN(q,u)$, $KN(q,u)$, and $SC(q,u)$ quantifying how well the query fits to user u from a social network/-knowledge/social capital perspective, the overall $score(q,u)$ is defined as

$$score(q,u) = \lambda_{SN} \cdot SN(q,u) + \lambda_{KN} \cdot KN(q,u) + \lambda_{SC} \cdot SC(q,u). \tag{29}$$

The optimal values for the parameters λ_{SN}, λ_{KN}, and λ_{SC} need to get ascertained empirically, e.g. using a training dataset and explicit relevance judgments from users. In the following, each component is explained in greater detail.

7.1.1 *Social Network*

As in existing literature (Wasserman and Faust, 1994), a social network is defined as a graph $G = (V, E)$ with a set of vertices V (the users) and a set of directed edges $E = V \times V$ (the relations between the users). Each edge stores a set of social attributes characterizing the social relationship between the two users who are connected by the respective edge. The set of attributes may contain tie strength, sympathy, similarity of social context, and similarity of content knowledge. Each attribute may be quantified with a ratio-scaled variable in the interval $[0, 100]$.

If the quality of a reply received in a social information retrieval scenario can get correlated to certain edge attributes, it would make sense to use this information to give more weight to those contacts who are connected to the information seeker with

an edge meeting the identified criteria. The objective is to verify to what degree those social attributes influence the performance of a social information retrieval system for a specific user.

Assuming that *tie strength* and *sympathy* have been identified as edge attributes that positively correlate with the relevance judgment of the user (as suggested by the results of Experiment 7a later, cf. Section 16.6.1.3), the result of $SN(q,u)$ could be defined as

$$SN(q,u) = (\mu_1 \cdot (\text{tie strength})(u) + \mu_2 \cdot (\text{sympathy})(u)) \cdot \left(\frac{1}{\sum_i \mu_i \cdot 100}\right) \quad (30)$$

with $(\text{tie strength})(u)$ and $(\text{sympathy})(u)$ representing the *tie strength* and *sympathy* values to user u, and μ_1 and μ_2 set to reasonable values (e.g., $\mu_1 = \frac{0.17}{0.17+0.22} = 0.44$ and $\mu_2 = \frac{0.22}{0.17+0.22} = 0.56$, cf. Table 22 in Section 16.6.1.3). To normalize all elements of the overall scoring function (Equation 29), $SN(u)$ gets multiplied with $\left(\frac{1}{\sum_i \mu_i \cdot 100}\right)$. The approach can be easily adopted to the quality of the information source to e.g., include only one social attribute type like tie strength for each edge. However, the concept requires the network and its attributes to be stated explicitly. This can be achieved in several ways, e.g.

- by using interaction logs, e.g. from email correspondence (Kossinets et al., 2008; Zhang and Ackerman, 2005), mobile communication networks (Saramäki and Onnela, 2007), or interaction on online social networks like Facebook (Panovich et al., 2012; Gilbert and Karahalios, 2009; Kahanda and Neville, 2009),
- by relying on proximity sensors (Aharony et al., 2011; Groh et al., 2010; Olguin and Pentland, 2009; Eagle and Pentland, 2006),
- by analyzing audio signals for turn-taking patterns in conversations (Groh et al., 2011c), or
- by assessing the data explicitly (like it was done in Experiment 7 in Chapter 16; most likely only feasible in experimental settings).

If user u is not a direct contact of the information seeker, egocentric approaches do not allow to derive values for the social attributes without a central perspective on the network. Therefore, the availability of an explicit social network as a backbone is assumed (e.g., like conventional online social networks), where users can explicitly publish their social edges (without the value of the social attributes for the edges). Visibility of the plain edges can be adjusted for certain groups, e.g., for friends only (as it is possible on today's online social network platforms as well). Using this structural information, it is possible to infer a hypothetical probability for tie strength to socially more distant users, e.g. as suggested in (Golder and Yardi, 2010). In absence of a real tie strength value, this indirect approximation could be used as an initial estimate for a measure representing social closeness.

Even without formal tie strength attributes, structural information (e.g., shortest path length, number of different shortest paths) could be used, depending on the network's available services.

7.1.2 *Knowledge Advertisement*

In the offline world, each person has an idea of what knowledge socially close people might have. This idea is fed by previous interaction with or chatter about the respective person. For a digital solution, a protocol to exchange information about available knowledge is required – users need to be able to *advertise* their own or other users' knowledge to make others aware of available information.

Such a system would have several requirements. A user must be able to adjust the advertised knowledge for each recipient. As in the offline world, users must be able to control their representation and adapt to the context of the relationship. While social closeness might be a good predictor for the amount, nature, and privacy sensitivity of advertised knowledge, it is not the only parameter determining it (e.g., there might exist cases where advertising knowledge only depending on social closeness might be a bad choice). The control mechanisms must allow to adjust the communicated knowledge at an acceptable granularity. Previous research shows that too complicated/granular privacy settings cause users to accept the application's default (Lipford et al., 2008). A peer-to-peer approach would be preferred: among other advantages, this would remove the requirement for a trusted third party and improve scalability.

In the following, the knowledge a user u advertises towards a user v is defined as the *knowledge profile* $\vec{k}_{u,v}$. Various archetypes of possible solutions for the advertising protocol can be pursued:

- *No Advertising*: A quite simple solution would be to avoid advertising at all. Instead of exchanging explicit knowledge profiles, users could speculate about other people's knowledge and build individual knowledge profiles of other users based on previous interaction and traditional meta information (occupation, age, gender, etc.). As a result, queries might have to be sent to a larger audience due to missing cues which would help finding the right recipient, especially when trying to exploit effects like serendipity.

- *Knowledge Profile Directory*: An approach based on the idea of "yellow pages" (Huang et al., 2013; Wagner et al., 2012; Yang et al., 2008), where a list of users with certain expertise can get retrieved, would require a personalized view on the directory, since each user needs to be able to customize her knowledge profile for each other user. Having multiple versions of each user's knowledge in the directory would be possible, as long as access is limited accordingly and it is ensured that users can not access knowledge profiles created for other users.

- *Distributed Knowledge Advertising*: Users could exchange information about the knowledge areas they would like to advertise in a distributed way: instead of relying on a central instance, users could inform each other proactively (e.g., actively send information about available knowledge) or reactively (e.g., send information about available knowledge upon request) and exchange knowledge profiles customized for the respective recipient.

While the first option appears to be a pragmatic and lightweight approach, it does not work efficiently in larger scenarios: to increase the probability of receiving a useful answer, an information seeker has to send the query to a large group of potential

information providers. Without a clear representation of other users' expertise, the required number of recipients for each query would be much higher than in any other system where explicit knowledge profiles are exchanged among the users. Especially when considering indirect contacts as well, the number of potential information providers would increase exponentially. In addition, this approach would harm the information seekers' privacy with regard to the information need and waste resources of the potential information providers, who would receive many information needs (without being able to answer the majority of them). However, the chances for serendipity might be higher than in approaches with knowledge profiles, since it is possible to identify a user who would not have listed the specific information in her knowledge profile. Options two and three would both be suited to serve the purpose, however, option three offers more flexibility and scalability. Some of the disadvantages of option 2 could get mitigated by encrypting the shared knowledge profiles with well-established asymmetric cryptographic tools. The concept as described in the following would work with both approaches.

To reflect a user's expertise to its full extent and to reduce the effort to create an expertise profile, the exchanged profiles are based on the user's information space (Chapter 8). Possible ways to reduce the dimensionality of the information space and calculate $\vec{k}_{u,v}$ include ESA and LDA: customizing the knowledge profile based on certain recurring semantic concepts (ESA) or latent topics (LDA) is a task that is possible (cf. Experiment 4 in Chapter 13), especially if only the most dominant areas are considered and additional clustering is used (e.g., by reducing the set of to central concepts in ESA or using fewer topics in LDA). In theory, TF-IDF (without partitioning in topics) is also possible, however, the dimensionality of the knowledge profile would most likely be too large to allow easy customization by the information provider (i.e., setting selected vector dimensions to 0). However, given its popularity, the approach is interesting for benchmark reasons (Experiment 4, Chapter 13). User u's convenience could be further increased by recommending a knowledge profile vector $\vec{k}_{u,v}$ (reflecting u's advertised knowledge towards v) based on the knowledge vectors u already customized for users similar to v (e.g., based on structural information of the social network). Especially for users with a higher distance on the social graph, the changes in the respective knowledge profile vector will most likely be negligible and converge to a "default" setting that is reserved for far-distant acquaintances. In the following, different ways to calculate knowledge vectors are presented. An empirical evaluation is done in Experiment 4 (Chapter 13).

EXPLICIT SEMANTIC ANALYSIS (ESA) Using Wikipedia as background corpus, a standardized set of concepts is available. As explained in Section 5.5, documents can get represented as vectors $\vec{d}$ in a vector space defined by concepts mentioned in Wikipedia ("Wikipedia articles"). Each value $\vec{d}_i$ represents the degree of relatedness of the document and the respective Wikipedia concept that constitute the dimension i. As in existing literature (Gabrilovich and Markovitch, 2009), TF-IDF can be used to calculate the degree of relatedness between a term (and in further instance, a document) and a Wikipedia article. A knowledge profile $\vec{k}_{u,v} = (k_1, k_2, ..., k_m)_{u,v}$ therefore is the ESA representation of u's information space, where each k_i represents the relatedness of u's information space to the concept represented by dimension i in the

vector space. User u's knowledge profile vector for v, $\vec{k}_{u,v}$, might have been adjusted by u to conceal certain areas of u's expertise to protect her privacy. When v has an information need, she calculates a query vector $\vec{q} = (q_1, q_2, ..., q_m)$, representing the relatedness of the query's textual representation to each of the m Wikipedia concepts defining the vector space. The query vector $\vec{q}$ and $\vec{k}_{u,v}$ can be compared using cosine similarity. Since all values in the vectors are positive, the result ranges from 0 (not similar) to 1 (identical) and therefore can directly be used as a definition for $KN(q, u)$ in Equation 29.

ESA-IDF In addition to the pure ESA approach explained above, the information seeker v considers the scarcity of certain expertise in her social network when evaluating the similarity of her query $\vec{q}$ and u's knowledge profile $\vec{k}_{u,v}$. Just like "Inverse Document Frequency" in TF-IDF (Section 2.3.2), v calculates for each $i \in \{1, 2, ..., m\}$ in the m-dimensional vector space of Wikipedia concepts for how many received knowledge profile vectors $k = (k_1, k_2, ..., k_m)$ the expression $k_i > 0$ holds. The results are stored in a frequency vector $f = (f_1, f_2, ..., f_m)$. Afterwards, each k_i of a knowledge profile vector $k = (k_1, k_2, ..., k_m)$ is multiplied with

$$\log\left(\frac{N_A}{f_i}\right). \tag{31}$$

The term N_A reflects the number of knowledge profile vectors in v's collection. For comparison, cosine similarity is used as in the plain ESA case.

ESA-LINK Apart from describing semantic concepts, Wikipedia articles are linked to each other. Each concept can be seen in a certain context of semantically close or related concepts, using Wikipedia's link structure as representation of semantic closeness among concepts. A user u can be a relevant information provider for an information need that is strongly related to concept c, even if u's information space does not reflect a relationship to c, but instead is strongly related to c' and c'' which both are strongly linked to c (links in both directions). This approach would also reduce the sparseness problem when comparing knowledge profile vectors, since the chance to find a positive match is much higher when each concept's neighborhood is included.

To compare a query vector $\vec{q} = (q_1, q_2, ..., q_m)$ and a knowledge profile vector $\vec{k} = (k_1, k_2, ..., k_m)$, the following formula may be used:

$$sim(\vec{q}, \vec{k}) = \sum_{i}^{n} \left(\sum_{j \in l(i)} \frac{q_i \cdot k_j}{d(i,j)+1} \cdot \frac{1}{\sqrt{\sum_i^n q_i^2} \cdot \sqrt{\sum_i^n k_i^2}} \right) \tag{32}$$

In the definition above, $d(i, j)$ expresses the distance (length of the shortest path) between the Wikipedia articles related to dimensions i and j in the graph defined by the links between Wikipedia articles. $d(i, j)$ is used to reduce the positive effect of existing expertise in related areas with increasing distance between the requested dimension and the available dimension. The function $l(i)$ defines the set of dimensions that are close to dimension i in the Wikipedia link structure, i.e. $x \in l(i) \Leftrightarrow d(i, x) \leqslant \epsilon$ for a parameter ϵ. Depending on the computational power of the device and the size

of the underlying concept space, ϵ needs to be adjusted accordingly. An additional question that has to be answered is whether incoming and outgoing links or only one type of links should be considered.

The formula in Equation 32 is based on the idea of cosine similarity. It can be seen as a generalization to smooth the information provider's knowledge space based on the link structure of the Wikipedia articles.

To normalize the result of Equation 32 on the interval $[0, 1]$ (and therefore allow to act as $KN(q, u)$ in Equation 29), the following small adjustment needs to be made: the inner sum needs to be multiplied by $\frac{1}{|l(i)|}$ with $|l(i)|$ representing the number of related concepts of i that are considered in the similarity calculation. Apart from that, no further adjustments are needed. The formula is based on the cosine similarity, which in turns ensures a result in the range $[0, 1]$ for positive input values. The only difference between plain cosine similarity and Equation 32 is that for each dimension of the query vector, we consider multiple dimensions in the knowledge profile vector as relevant. Since this effect is taken care of by the adjustment explained above, the definition mentioned in Equation 33 can be a definition for $KN(q, u)$ in Equation 29.

$$\text{sim}(\vec{q}, \vec{k}) = \sum_{i}^{n} \left(\frac{1}{|l(i)|} \sum_{j \in l(i)} \frac{q_i \cdot k_j}{d(i,j)+1} \cdot \frac{1}{\sqrt{\sum_i^n q_i^2} \cdot \sqrt{\sum_i^n k_i^2}} \right) \tag{33}$$

LDA Each author calculates a set of topics based on her information space using LDA (Section 5.4.1). Since each latent topic is defined as a probability distribution over words, the knowledge profile of a user u is a list of distributions over words. Before sharing the knowledge profile, u can adjust it by removing topics she does not want others to know about. This approach is the more generalized variant of a tagging system: a tag can be seen as a topic T with only one term t (and the probability distribution for the topic being defined as $P(X = t|T) = 1$). Therefore, it is also possible for u to manually add "topics".

An information seeker v with a set of knowledge profiles from other users $u \in U$ identifies the set of designated information providers by conducting the following steps:

- For each user u, who shared her knowledge profile with v:
  - Calculate the Jensen-Shannon distance (Section 5.4.2) for the query and each topic of u's knowledge profile (using the same preprocessing algorithms, e.g. like stemming)
  - Calculate the average distance for the user (adding the distances for each topic and dividing by the number of topics)
- Choose the users with the highest average value

Following this approach, there might be a bias in the evaluation of users with a large number of clearly separated topics (few overlaps of terms) when compared to users with less topics and/or less separated topics. In the first case, a query containing only a term t would have a positive value in a small set of topics, leading to a small average value for this user because of the large divisor. In the second case, t

would be part of many topics (and therefore might lead to a higher numerator). If the number of topics is low, the small divisor might cause a higher average distance. In case empirical evidence would support this assumption, an alternative to the mean value for aggregation on user level would be to use the minimum distance instead.

Depending on the base of the logarithm used to calculate Jensen-Shannon divergence, Jensen-Shannon distance is bounded by $\sqrt{0}$ and $\sqrt{1}$ (base 2) or $\sqrt{0}$ and $\sqrt{\log_e(2)}$ (base e). In the latter case, the result needs to be divided by the upper bound to be normalized on the interval $[0, 1]$. Since the Jensen-Shannon distance represents a distance (and not a similarity measure), the result needs to be converted to a similarity measure by subtracting it from 1 before being used as replacement for $KN(q, u)$ in Equation 29.

7.1.3 *Market Model for Social Capital*

7.1.3.1 *Motivation*

Sharing information can be interpreted as an activity taking place in a special kind of market, which has parallels to conventional economic market models (though, there also are differences, e.g. it is socially not fully accepted among friends or acquaintances to actively negotiate or bargain about the price for an information item). In Section 3.1, existing ideas for market approaches have been briefly discussed. The concept that would be most applicable for the distributed social information retrieval scenario introduced in this thesis is the market model presented by Groh and Birnkammerer (Groh and Birnkammerer, 2011; Birnkammerer, 2010) for the following reasons:

- It provides an explicit representation of capital flows, is based on mathematical equations, and can get operationalized directly.
- It is targeted to the exchange of information in a social context and therefore covers mechanisms like variable costs due to loss of privacy or the respective roles (information provider / information seeker).

However, while the model offers valuable contributions and ideas, it does not fully cover the scenario of the proposed social information retrieval concept. Its unit of measure is the single transaction of information item i from information provider A to information seeker V, written as $T_i^{A \to V}$. It is not possible to e.g. rely on social goodwill by sharing information without an immediate compensation, which might arise at a later time or might be part of an additional information exchange that is skewed in the other direction (by allowing different marginal income levels, where e.g. information i generates a high marginal income and therefore compensates for information item j which has been given away for a low or even non-existent marginal income). In the following, Groh and Birnkammerer's model will get expanded to better match the intended domain of this thesis. In the extended version of this model, three fundamental additions are made:

- *Decline over time*: (Social) capital flows are considered in their respective time period where they manifest themselves. As in financial modeling (Berk and

DeMarzo, 2014, p. 63), where money is given a time value, social capital should also be linked to time. A favor that has been received a long time ago (or that is promised to be done in the future) is not as valuable as a recently gained favor of the same "utility class".

- *Consideration of Social Environment*: When deciding to help or ask someone for help, the respective social environment is explicitly taken into account. Relying on an already established contact who connects two previously unknown users helps to form a (temporal) social group, which might benefit from the ingroup effect (Tajfel, 1970).
- *Modeling of Information Seeker* The information seeker's decision is also part of the model: markets are not one-sided and therefore should cover both parties.

The specific implementation of the extensions mentioned above is not the only possible way of reflecting the outlined ideas. Due to limited related work, the following model should be seen as a first attempt to formalize human information sharing behavior as exchanges on an information market. The model's basic components are two equations: the first one (Equation 36) supports the information provider when deciding whether or not to share an information item with an information seeker and the second one (Equation 53) facilitates the information seeker's decision whether or not to ask a specific person for information.

7.1.3.2 *Decision Function for Information Providers*

Let A be a potential information provider who received a request from an information seeker, V, for a specific information item. Assuming A acts rationally, she shares the desired information item iff A's balance sheet for transactions with V does not turn too negative by doing it, i.e.

- A either had a positive net gain when acting as an information provider towards V (difference between gain $G^{IP}_{A,V}$ and respective costs $C^{IP}_{A,V}$ is positive) and/or
- A's net gain from interacting as an information seeker with V (difference between gain $G^{IS}_{A,V}$ and respective costs $C^{IS}_{A,V}$) is positive and compensates a potential negative contribution of the previous point.

When writing both requirements in an equation, the result would be the following:

$$(G^{IP}_{A,V} - C^{IP}_{A,V}) + (G^{IS}_{A,V} - C^{IS}_{A,V}) \geqslant 0 \Leftrightarrow \tag{34}$$

$$G^{IS}_{A,V} - C^{IS}_{A,V} \geqslant C^{IP}_{A,V} - G^{IP}_{A,V} \tag{35}$$

Equation 34 suggests that the net gain from the interactions where A acted as an information seeker should be higher than the costs caused by sharing information with V (reduced by the gains that could be realized by sharing the information).

To facilitate market transactions, it sounds plausible to not reject a request in borderline situation but to concede *goodwill* $GW_{A,V}$ to the requesting party. This goodwill can be interpreted as a loan the information provider A is willing to give to the information seeker V:

$$(G^{IP}_{A,V} - C^{IP}_{A,V}) + (G^{IS}_{A,V} - C^{IS}_{A,V}) + GW_{A,V} \geqslant 0 \Leftrightarrow \tag{36}$$

$$G^{IP}_{A,V} + G^{IS}_{A,V} + GW_{A,V} \geqslant C^{IP}_{A,V} + C^{IS}_{A,V} \tag{37}$$

If the inequation is true, A should help V, otherwise it is not rational for A to do so. The reaction in such a case are discussed in Section 7.1.3.5.

In the equation above,

- $G^{IP}_{A,V}$ refers to the GAIN A received when interacting as *information provider* with V (and those gains that can get attributed to V), e.g. A gains additional popularity, trust, or social status when sharing information with V.
- $G^{IS}_{A,V}$ refers to the GAIN A received when interacting as *information seeker* with V (and those gains that can get attributed to V), e.g. quality and utility of information provided by V
- $C^{IP}_{A,V}$ refers to the COSTS A had to bear when interacting as *information provider* with V (and those costs that can get attributed to V), e.g. costs in assembling the information or loss of privacy.
- $C^{IS}_{A,V}$ refers to the COSTS A had to bear when interacting as *information seeker* with V (and those costs that can get attributed to V), e.g. the price A had to pay (time to consume and interpret the information, initial socials costs caused by sending the query).
- $GW_{A,V}$ represents the GOODWILL A would have towards V, i.e. the quantified concession (which depends on the individual personalities of A and V and the type of relationship between A and V).

Throughout the model, the term GAIN is used as the additional benefit a person receives by a certain action – the costs to achieve those benefits are not subtracted yet. In contrast, NET GAIN refers to net benefit (i.e., GAIN reduced by the COSTS to generate the gain). In the following, all components of Equation 36 will be explained in detail.

A'S GAIN WHEN ACTING AS INFORMATION PROVIDER A's gain when acting as an information provider towards V can get expressed as

$$G^{IP}_{A,V} = \bar{G}^{IP}_{A,V} + \sum_{W \in S(V,\alpha)} \left(\bar{G}^{IP}_{A,W} \cdot \frac{1}{(1+\beta)^{d(V,W)}} \right) \tag{38}$$

where $S(V, \alpha)$ describes the set of users that are socially close to V (with α being the distance threshold) according to V's social graph (as assumed by A), β being the discount factor for socially close contacts and $d(V, W)$ represents the social distance between V and W. $\bar{G}^{IP}_{A,V}$ and $\bar{G}^{IP}_{A,W}$ represent the actual gain functions directly following from (Groh and Birnkammerer, 2011; Birnkammerer, 2010) and explained below in Equation 39. Considering the additional gain from socially close peers of V appears to be useful, when assuming that a positive attitude of one group member

towards A (in this case, V) might influence others ($S(V, \alpha)$). The fact that people influence their peers has been measured in various ways before and constitutes an own area of research (Rao et al., 2015; Tinati et al., 2012).

$$\bar{G}^{IP}_{A,V} = \sum_{x \in T^{A \to V}} \left(P(x) + R(x) + \sum_{y \in \Gamma(x)} R(y) \right) \tag{39}$$

$T^{A \to V}$ is a generalization of Groh and Birnkammerer's notation and describes the set of all information transactions from A to V. $\Gamma(x)$ is defined as in (Birnkammerer, 2010) and represents the set of users that received the information item from V and also increased A's reward. PRICE $P(x)$ and REWARD $R(x)$ refer to the same concepts as originally introduced in (Birnkammerer, 2010), however, to include depreciation in the model, an approach based on Net Present Value (NPV) is used to depreciate transactions based on their age:

$$R(x) = \sum_{r} \frac{r}{(1+\gamma)^{-t(r)}} \tag{40}$$

$$P(x) = \sum_{p} \frac{p}{(1+\gamma)^{-t(p)}} \tag{41}$$

In Equation 40 and Equation 41, REWARD R and PRICE P of a transaction are defined as a set of individual "*cash flows*" (referred to as r and p), which can manifest themselves at different points in time. $t(r)$ and $t(p)$ represent the temporal distance between the current point in time and the situation in which r or p occurred. Due to the backward-looking architecture, $t(.)$ returns negative values, therefore the result is considered as $-t(.)$ in Equation 40 and Equation 41. The parameter γ defines the rate of depreciation for each time period.

To summarize, $G^{IP}_{A,V}$ contains the gain that A received when acting as an information provider for V, considering

- the gain obtained from V during all previous information transactions from A to V (to reflect the fact that social interactions develop iteratively (Nowak, 2006)),
- the gain obtained from people within V's social network where the additionally received reward can be ascribed to V (relevant group is defined with threshold α and discounted with β), and
- a discounting effect for aged information (discount rate modeled with γ), considering asynchronous release of reward and/or price.

The current information item V is requesting and its future rewards are considered in Equation 38 using estimates.

A'S GAIN WHEN ACTING AS INFORMATION SEEKER Following the same logic as above, A's gain when acting as Information Seeker can get expressed as

$$G^{IS}_{A,V} = \bar{G}^{IS}_{A,V} + \sum_{W \in S(V,\alpha)} \left(\bar{G}^{IS}_{A,W} \cdot \frac{1}{(1+\beta)^{d(V,W)}} \right) \tag{42}$$

with

$$\bar{G}^{IS}_{A,V} = \sum_{x \in T^{V \to A}} \left(U(x) \cdot \frac{1}{(1+\gamma)^{-t(x)}} \right) \tag{43}$$

and $U(x)$ representing the utility user A has when consuming the information transferred in transaction x from V to A, discounted with γ ($t(x)$ again returns the distance between now and the period when $U(x)$ was realized).

Again, the idea is that the relationship between A and V is not only characterized by their individual previous transactions (where A asked for information from V) but also by the interactions A had with V's closer circle of friends.

A'S COST WHEN ACTING AS INFORMATION PROVIDER When acting as an information provider, A has to bear costs when sharing information with V. The concept of fixed costs $K_f(x)$ and variable costs $k_v(x) = k^T_v(x) + k^P_v(x)$ for a transaction x are considered as defined in the original approach (cf. (Groh and Birnkammerer, 2011; Birnkammerer, 2010) and Section 3.1). However, in contrast to the original proposal, the approach presented here does also discount costs that arose in previous time periods and considers costs created by sharing information with socially close people:

$$C^{IP}_{A,V} = \bar{C}^{IP}_{A,V} + \sum_{W \in S(V,\alpha)} \left(\bar{C}^{IP}_{A,W} \cdot \frac{1}{(1+\beta)^{d(V,W)}} \right) \tag{44}$$

with

$$\bar{C}^{IP}_{A,V} = \sum_{x \in T^{A \to V}} \left(K_f(x) + k_v(x) + \sum_{j \in \Gamma(x)} k^p_v(x) \right). \tag{45}$$

Like in the previous cases, $S(V,\alpha)$ refers to the set of users who are – according to A's assumption – part of V's social network, $d(.)$ is the social distance function introduced above and β is again the discounting factor for social contacts of V. As mentioned already above (done with $P(x)$ and $R(x)$, $K_f(x)$, $k_v(x) = k^T_v(x) + k^P_v(x)$ are discounted with a discounting factor γ to reduce the effect of costs that have been realized in previous time periods:

$$K_f(x) = \sum_f f \cdot \frac{1}{(1+\gamma)^{-t(f)}} \tag{46}$$

$$k_v^T(x) = \sum_{v^T} v^T \cdot \frac{1}{(1+\gamma)^{-t(v^T)}} \tag{47}$$

$$k_v^P(x) = \sum_{v^P} v^P \cdot \frac{1}{(1+\gamma)^{-t(v^P)}} \tag{48}$$

As in the case of $G^{IP}_{A,V}$, $t(.)$ acts as a temporal distance measurement function and γ is the discount factor for previous costs.

A'S COST WHEN ACTING AS INFORMATION SEEKER As in the previously explained costs $C^{IP}_{A,V}$, A's cost when interacting with V as an information seeker can be split in the costs that can be directly associated with V and the second (and with β discounted) part, which consists of users socially close to V (using α as a threshold for the definition of the social network). It is defined as

$$C^{IS}_{A,V} = \bar{C}^{IS}_{A,V} + \sum_{W \in S(V,\alpha)} \left(\bar{C}^{IS}_{A,W} \cdot \frac{1}{(1+\beta)^{d(V,W)}} \right) \tag{49}$$

with the same function names and parameters as in the previous cases. The cost function covers the price paid for the received information and the costs of initiating the request and is defined formally as

$$\bar{C}^{IS}_{A,V} = \left(\sum_{x \in T^{V \to A}} \frac{P(x)}{(1+\gamma)^{-t(x)}} \right) + \left(\sum_{x \in RQ^{A \to C}} \frac{AC(x)}{(1+\gamma)^{-t(x)}} \right) \tag{50}$$

with $P(x)$ being the price A paid for each received information from V, γ being the discount factor to reflect depreciation over time, $RQ^{A \to V}$ being the set of requests sent from A to V and $AC(.)$ quantifying the social costs of initiating the interaction. To allow depreciation over time, it is multiplied with $\frac{1}{(1+\gamma)^{-t(x)}}$.

7.1.3.3 *Decision Function for Information Seekers*

Two criteria are relevant for the decision of information seeker A whether or not to send a request to a social contact V:

1. How large is the expected net gain of the answer, i.e., how big is $G^{IS}_{A,V} - C^{IS}_{A,V}$? A positive net gain makes asking V attractive.
2. How much effort did the information seeker A invest to help V? The higher the costs A had to bear when supporting V, the more legitimate it is to ask V for a favor in return. A's costs when helping V (i.e., sharing information with V) are expressed as $C^{IP}_{A,V}$, those costs need to get reduced by the gain A received from V for sharing that information. It is important not to fully deduct $G^{IP}_{A,V}$

from $C^{IP}_{A,V}$ because this would include potential rewards caused (and payed for) by other users to V's account. Therefore, $C^{IP}_{A,V}$ gets only lowered by $G^{IP}_{A,V}|V$ which can get interpreted as the gain A received for sharing information with V that is directly payed for by V.

Summarizing both requirements in an equation would lead to

$$(G^{IS}_{A,V} - C^{IS}_{A,V}) + C^{IP}_{A,V} - G^{IP}_{A,V}|V \geqslant 0 \tag{51}$$

$$G^{IS}_{A,V} - C^{IS}_{A,V} \geqslant G^{IP}_{A,V}|V - C^{IP}_{A,V}. \tag{52}$$

Equation 51 suggests that the net gain from the interactions as information seeker should be higher than the negative effect of the costs when acting as information provider – or stating it differently, bearing high costs for helping V can justify asking V for help, even if V's responses have not been of high value in the past or are only available at high costs.

In Equation 52, it is possible to identify an interesting effect: a very high value of $G^{IP}_{A,V}|V$ could cause an information seeker A not to ask V. If the price V had to pay for A's help was exceptionally high, the right side of Equation 52 would be greater than the left side. This is the main reason why $G^{IP}_{A,V}|V$ is used instead of $G^{IP}_{A,V}$. It could be seen as realistic in certain circumstances that A, who could realize a huge net gain by sharing information with V, would not dare to ask V for another favor. To also allow the opposite behavior and to have an equivalent to *goodwill* $GW_{A,V}$ in the information provider's equation explained above, the decision equation for information seekers is extended with an additional *audacity* factor $AD_{A,V}$ which models A's boldness towards V:

$$(G^{IS}_{A,V} - C^{IS}_{A,V}) + C^{IP}_{A,V} - G^{IP}_{A,V}|V + AD_{A,V} \geqslant 0 \tag{53}$$

$$G^{IS}_{A,V} - G^{IP}_{A,V}|V + AD_{A,V} \geqslant C^{IS}_{A,V} - C^{IP}_{A,V} \tag{54}$$

$AD_{A,V}$ depends like $GW_{A,V}$ on the personalities of A and V and on the type of social relation both parties have. Like $GW_{A,V}$, it also facilitates transactions by defining a broader margin where it is rational for A to send a request to V.

$G^{IP}_{A,V}|V$ refers to the gain A received directly from V when interacting as information provider with V (i.e., the part of $G^{IP}_{A,V}$ which creates direct costs for V). Following the definition of $G^{IP}_{A,V}$ and $C^{IP}_{A,V}$, $G^{IP}_{A,V}|V$ could be defined analogously:

$$G^{IP}_{A,V}|V = \bar{G}^{IP}_{A,V}|V + \sum_{W \in S(V,\alpha)} \bar{G}^{IP}_{A,W}|W \cdot \frac{1}{(1+\beta)^{d(V,W)}} \text{ and} \tag{55}$$

$$\bar{G}^{IP}_{A,V}|V = \sum_{x \in T^{A \to V}} P(x) \tag{56}$$

7.1.3.4 *Overview of Variables and Functions*

As a quick reference, Table 4 contains a summary of the variables introduced in the previous two sections.

VARIABLE	DESCRIPTION
α	Parameter for $S(.)$ function to limit the number of social contacts of the interaction partner to be considered when calculating $G(.)$ and $C(.)$
β	Depreciation rate for social relationships – the more distant a social contact is, the less influence does she have on the overall decision
γ	Depreciation rate for past interactions – events that took place long ago in the past do not impact the overall decision to the same extent as current events
$G_{A,B}^{Mode}$	Overall gain for user A when interacting in role *Mode* with user B
$C_{A,B}^{Mode}$	Overall costs for user A when interacting in role *Mode* with user B
$GW_{A,B}$	Goodwill, i.e. concession A is comfortable to making towards B
$AD_{A,B}$	Audacity, i.e. concession A is assuming to receive from B (also interpreted as A's boldness when contacting B)
$\bar{G}_{A,B}^{Mode}$	Gain for user A when interacting in role *Mode* with user B (not considering social network of B)
$\bar{C}_{A,B}^{Mode}$	Costs for user A when interacting in role *Mode* with user B (not considering social network of B)
$d(A,B)$	Social distance between two users A and B, measured e.g. in tie strength or hops
$S(A,\alpha)$	Set of other users who have at most a distance of α to A, i.e. $X \in S(A,\alpha) \Leftrightarrow d(A,X) \leqslant \alpha$
$T^{A\to B}$	Set of transactions from A to B
$U(x)$	Utility of transaction x for the receiver of the information
$R(x)$	Reward for the information provider for conducting transaction x
$P(x)$	Price paid for transaction x by the receiver of the information (money, time to consume the information, explicitly agreed terms between information provider, etc.)
$K_f(x)$	Fixed costs for information provider to create/get information transferred in transaction x
$k_v(x)$	Variable costs for information provider to transfer information in transaction x – can be split into $k_v^P(x)$ and $k_v^T(x)$
$k_v^P(x)$	Part of variable costs $k_v(x)$ reflecting the costs of the loss of privacy for the information item transferred in transaction x
$k_v^T(x)$	Part of variable costs $k_v(x)$ reflecting the technical costs of the transfer x
$t(x)$	Temporal distance between now and the point in time when event x (transaction, payout of reward/price/utility) took place or will take place
$\Gamma(x)$	For a transaction $x \in T^{A\to B}$, $\Gamma(x)$ is the set of users who received x's content from B, i.e. $X \in \Gamma(x) \Leftrightarrow \exists y \in T^{B\to X} \wedge inf(x) = inf(y)$ with $inf(x)$ representing the information transferred in information transaction x
$G_{A,B}^{IP}\|B$	Gain for user A directly received from B (and/or B's social network) when interacting as information provider with user B

Table 4: Variables in market model explaining social capital flows

7.1.3.5 *Relevance for Reality*

Depending on the type of the social contact, information exchanges follow different patterns. As already mentioned in (Adler and Kwon, 2002, p. 34), the characteristics of a market encounter highly depend on the participants' social relationship (cf. Section 3.1). Following Fiske's classification for interaction types introduced in Section 2.4.2 and summarizing Fiske's *Equality Matching* (EM) and *Market Pricing* (MP) in Adler and Kwon's *Market Relations* category suggests that those social relations have specific terms of exchange and an explicit agreement on the exchange. The exchange can also be considered symmetrical. Examples include interactions between strangers. In *Hierarchical Relations* (matching Fiske's *Authority Ranking* regime), the terms of exchange are diffuse (e.g., employment contracts do not list exactly all duties of an employee), however, the terms of exchange are explicit (e.g., it is transparent for all involved parties that there is an employment contract, even if it states the obligations of the employee only diffuse). Furthermore, the exchange is asymmetrical. For *Social Relations* (vaguely corresponding with Fiske's *Communal Sharing* regime), the terms of the exchange are diffuse (a favor done today is made for exchange with a favor to be done at some undefined point in time in the future) and the terms of the exchange are tacit (i.e., a favor is done in the understanding that it will get returned someday in the future). The relationship can be considered as symmetrical (Adler and Kwon, 2002; Fiske, 1992).

In the following, the Social Capital Model for Information Exchange introduced above will be discussed with regard to Fiske's forms of sociality (treating EM and MP as one category). The objective of this section is to illustrate how the parameters of the model change if two participants would interact in a CS, AR, or EM/MP regime.

INTERPRETING THE MODEL WITH REGARD TO CS-HEAVY INTERACTIONS A social relationship that has a large CS component would typically follow a very altruistic behavior (cf. Section 2.4.2). This would generally result in higher values for goodwill $GW_{A,V}$ (the "giving" member would increase the maximum acceptable disbalance of the relationship). In addition, the costs to initially ask a potential information provider could be lower because people are more willing to help other members of the group (i.e., $C^{IS}_{A,V}$ would be lower). Furthermore, audacity $AD_{A,V}$ could be lower than in other scenarios (except for free riders) when people act altruistically.

INTERPRETING THE MODEL WITH REGARD TO AR-HEAVY INTERACTIONS For relationships that have a large AR component, one party has a higher authority than the other. This asymmetry will cause the initial costs of sending the query to the other party to change – the user with the higher rank will have less costs asking the user with the lower rank, while the costs will increase for the user with the lower rank. Depending on the individual personality, the user with the lower rank might increase her goodwill GW towards the higher-ranked user, assuming that the latter will return a granted benefit with a higher likelihood. With regard to audacity, a higher-ranked user A might have greater values for $AD_{A,V}$, allowing her to demand more from V – however, this also depends on A's personality. To a certain degree, the higher rank would give A a justification for increased levels of $AD_{A,V}$.

INTERPRETING THE MODEL WITH REGARD TO EM/MP-HEAVY INTERACTIONS Relationships with a strong EM/MP dimension are characterized by following rational, economic considerations – depending on individual type of social relationship and personal traits, the goodwill and audacity parameters will be adjusted to increase the scope for temporarily uneven relationships. In other situations, it might also get reduced (e.g., a rare interaction with a socially very distant user). It could be the case that goodwill and audacity depend on each other, both growing with increasing strength of the relationship.

RELEVANCE OF THE PROPOSED MODEL In typical social interactions, people will not act fully following the proposed model. The model's purpose is not to directly model users' immediate behavior, but to explain underlying relations. Social interactions always also follow a certain protocol, at least as long as the social interaction is not meant to terminate the social relation *abruptly* (for a "normal" termination of a social relation, common protocols exist). For example, in certain social settings it appears very unlikely that people bargain about the value and exchange terms for a specific information item – it would damage the relationship. Instead, the information provider would either accept a higher goodwill (and therefore make concessions and believe that the information seeker would even out the relationship) or neglect the availability of the information (and therefore avoiding any discussions about the price).

7.1.3.6 *Evaluation Concept*

The proposed market model could be evaluated using the following two-step evaluation process:

1. Scenario and/or interview based evaluation of the model's basic coherence
2. Long-term real-life experiment using mobile applications to predict social transactions

In the first phase of the evaluation, the objective is to verify the main ideas of the market model. This could be done using interviews or surveys and predefined scenarios. Participants could mention factors influencing their individual decision (to share an information item or to send a request to someone) in an open interview. Later, several predefined specific scenarios could be presented to the participants and the participants could assess how they would interpret the described situation in terms of concepts related to the market model.

After the first phase of the evaluation, the results should get screened thoroughly and checked for consistency with the model. In case of any identified needs for adjustment, the model should get adjusted.

In the second phase, the model could get implemented as a mobile application. The participants of the second phase should install it on their mobile device and manually enter each social transaction in the application (also those interactions that got rejected by the information provider). By using the application, each participant collects a set of social transactions and their individual response to them, which could be compared to the model's prediction of the response. It might make sense to

ask the participant to comment on some of the social interactions to allow an earlier adjustment of the model's parameters.

Due to the fact that the thesis' focus lies on different aspects, the evaluation steps as described above are left for future work. However, the model and the data obtained from Experiment 7 (cf. Chapter 16) are used to discuss the findings with regard to the research questions of this thesis in Section 16.6.1.5.

7.1.3.7 *Mapping Results for Scoring Function*

In order to be used as $SC(q, u)$ function from Equation 29, the result of the information seeker's decision function introduced in Section 7.1.3.3 needs to get transformed to the interval $[0, 1]$. An easy, straightforward way of doing this would be a binary response (0 or 1), depending whether the information seeker's decision function (Equation 53) is true (1) or not (0). While being pragmatic, this mapping does not pass information about the magnitude of the differences to the overall scoring function. This could be achieved with a slightly more complicated approach as follows: since the components of Equation 54 are not bound, transforming the outcome to an ordinal scaled result would require to calculate both sides of the equation for each possible information provider and scale the difference (left side minus right side) from 0 to 1 for results where the subtraction yields a positive result (and set it to 0 for all those users where the result is negative).

7.2 SOCIAL INTERACTION

7.2.1 *Query Modes*

The query mode is an attribute of the query that is set by the information seeker and differentiates between two structurally different modes: the first one deals with discrete information and uses the system to collect (and possibly aggregate) semantically tagged information ("semantic mode"). A use case one could think of is to get an overview of items chosen by friends (e.g., bought products, visited places, etc.). The result items provided by the information providers are all of the same information type and it therefore is possible to aggregate the findings. The second mode is more general and resembles the normal, non-semantic search for documents ("general purpose mode"). The information seeker sends a query and receives responses from the information providers in different formats (e.g., documents, links to resources, manually written information). Given the unstructured character of the data, it is not possible to automatically interpret the replies – the information seeker has to manually read the results to make sense of them. The main reasons to differentiate between the two modes are the following:

- *Adjust searched information space by focusing on relevant types of information items*: When an information seeker has an information need that requires a specific type of information item as response, it is useful to filter the searched information space of the information providers for that type to avoid noise (which might occur in general purpose search scenarios which rely only on the content of the information item). The reduction does remove the necessity to infer the

characteristics of the information item: does a document mentioning a certain product really imply that the information provider is interested in that product or is it just an advertisement? This vagueness of the reasoning can be removed by relying on predefined semantic structures.

- *Improve functionality by allowing interpretation when possible*: Relying on a fixed type of information items can help to interpret the results automatically. It becomes possible to aggregate and/or cluster information based on explicit inherent and semantically useful features of the information items (e.g., for restaurants: the type of cuisine, price range, distance). Such a chain of reasoning would be very difficult to achieve using general purpose search methods, relying on statistical methods without explicitly modeled concepts.

Apart from the advantages listed above, semantic concepts restrict the use cases, since they rely on predefined concept ontologies and require a data extraction and matching process considering these semantic structures. To allow unrestricted search, e.g. to identify expertise or perform a broad exploratory search in the social network, the general purpose mode is more versatile.

7.2.2 *Privacy Settings*

7.2.2.1 *Motivation*

Social search involves two parties, both reveal information about themselves, and both have the urge to protect their privacy when needed.

INFORMATION SEEKER Sending a query to a traditional search engine is an act that maintains the information seeker's privacy: apart from the internet service provider, the company operating the search engine and the operators of the visited websites, nobody can link the query to the information seeker. When using encrypted connections (HTTPS), no party has enough information to link the information seeker to the query. This protection is not in place when information seekers send their queries to members of their social network. It therefore needs to be possible to stay anonymous as an information seeker.

INFORMATION PROVIDER One could think of scenarios where an information provider does not want the information seeker to know that she gave a specific answer, even if it might be helpful (possible examples include e.g. information about illnesses and the individual experience with treatments).

Although it is possible to think of scenarios where anonymity of information seeker and provider are useful, it is only considered as last resort: some of the benefits of social search lie in the relationship between the information seeker and provider, which are only partially considered when one or both parties stay anonymous. The focus of this thesis is to investigate limits and chances of social information retrieval; therefore, the objective is to show under which constraints social information retrieval is useful and which problems and challenges occur. As a proof of concept, a privacy-preserving protocol for different scenarios is presented in the following. However, the protocol should be considered only as evidence that it is technically possible to

provide a system with the desired privacy-preserving functionality and not as the best possible solution to the problem.

7.2.2.2 *Privacy-Preserving Protocol for Social Information Retrieval (Proof of Concept)*

The protocol combines Chaum's mix cascade (Chaum, 1981) with Yang et al.'s approach for data collection (Yang et al., 2005). In the following, solutions for the scenarios where only the information seeker (1), only the information provider (2), or both parties (3) stay anonymous are presented. For each scenario, both communication paths (query: information seeker to information providers; response: information provider to information seeker) are modeled.

ANONYMOUS INFORMATION SEEKER The information seeker follows the concept of a mix cascade, described in (Chaum, 1981) and used e.g. in the Tor project[1]. The basic idea is that the information seeker encrypts her message to the information provider with the information provider's public key of an asymmetric encryption system like RSA (Rivest et al., 1978). In the next step, the information seeker defines a vector of n mixes $M = (m_1, ..., m_n)$ the message has to pass before it is delivered to the information provider. Each mix m_i is a computer system that receives a message, decrypts it with its private key $k_i^{private}$, interprets the forwarding instructions and executes them, i.e. sends the payload of the message to the next mix m_{i+1}. The forward instruction for the last mix, m_n, is to send the payload to the information provider. Each mix can only read its individual forwarding instructions, because the payload for the next layer is encrypted with the public key of the recipient of the next level. With $enc(c, k)$ being the encryption function of content c with public key k and $n = 2$, the information seeker sends the following information to the first mix, m_1:

$$enc\left(enc\left(enc\left(c, k_{ip}^{public}\right), k_2^{public}\right), k_1^{public}\right). \tag{57}$$

The first mix decrypts the message using the private key $k_1^{private}$, leading to

$$enc\left(enc\left(c, k_{ip}^{public}\right), k_2^{public}\right). \tag{58}$$

The process continues until the information provider receives

$$enc\left(c, k_{ip}^{public}\right) \tag{59}$$

and is able to decrypt it and read c using $k_{ip}^{private}$. The payload, c, contains the query and information how to reply. This includes a public key to encrypt the response, an address of a relay where the response should be sent to, and a session identifier for the interaction to distinguish different sessions. The information provider replies to the query by encrypting the response with the provided public key and sends it (combined with the session identifier) directly to the specified relay. The relay has been instructed by the information seeker to forward any incoming message from

1 https://www.torproject.org/ (retrieved 2015-12-13)

one of the recipients with the respective session identifier to the information seeker (without revealing the information seeker's identity to the information providers). It could be possible that each agent could act as a relay or that such systems are operated by trusted organizations. The information seeker receives the reply, and decrypts it to read the plain text.

ANONYMOUS INFORMATION PROVIDER For the scenario with anonymous information providers, a process based on Yang et al.'s approach is used (Yang et al., 2005). The basic idea behind the concept is that a set of public keys k_i^{public} can be combined to a key $k^{public} = \sum_i k_i^{public}$, which can be used to encrypt a message. Such an encrypted message can be decrypted successively using the individual private keys $k_i^{private}$ of the information providers. This allows every information provider to encrypt the reply with a public key of the group, which does not have a single private key equivalent. Instead, each information provider must decrypt the complete response of all information providers with her own private key consecutively (and while doing it, shuffles the order of the responses). In the following, the process steps are explained, for a detailed mathematical explanation, please refer to (Yang et al., 2005). The information seeker sends the query directly to the information providers $IP = \{ip_1, ip_2, ..., ip_n\}$. Each information provider ip_i is aware of the other recipients of the query and has published a public key k_i^{public} in a public key infrastructure available to everyone. Each ip_i retrieves the individual public keys for all other recipients of the query, i.e. for all members of the set IP, and calculates a combined public key, $k_{IP}^{public} = \sum_{j \in IP} k_j^{public}$. Each information provider ip_i encrypts her response r_i to the query with k_{IP}^{public}. Afterwards, no single member of the set IP can decrypt it alone. All information providers send $enc(r_i, k_{IP}^{public})$ to the information seeker. After each information provider submitted her data to the information seeker, the information seeker received $(enc(r_1, k_{IP}^{public}), enc(r_2, k_{IP}^{public}), ..., enc(r_n, k_{IP}^{public}))$. The information seeker sends this vector to one of the information providers, who (1) shuffles the responses by applying a random permutation and (2) uses her private key to decrypt the responses. The results are not (yet) in plaintext, because the other information providers must still use their keys successively to decrypt it. The information provider sends the message back to the information seeker, who sends the response vector to the next information provider. After the last information provider applied her private key (and again shuffled the responses), the responses are decrypted (and the original ordering can not be reconciled). In order to avoid that the last information provider in the chain can read the responses, it is possible to include the information seeker's public key in k_{IP}^{public}, so that she would be the one who performs the last decryption step.

ANONYMOUS INFORMATION SEEKER AND ANONYMOUS INFORMATION PROVIDER When both parties should act anonymously, the two approaches explained above need to be combined: the relay user who receives the reply of the information providers would step in and act as data aggregator and orchestrate the permutation and decryption cycles as explained above to shuffle the responses. Afterwards, she would send the results to the original information seeker.

7.2.2.3 *Discussion and Limitations*

The system assumes that the mixes act as they should and do not keep any information of forwarded packages. In addition, the relay agents who establish the link between the information providers and the anonymous information seeker need to be trustworthy. In the scenario with anonymous information providers, each information provider must know the complete set of information providers who received the request to "hide" within this group. Therefore, the group must be sufficiently large. In addition, the process does not work if $n-1$ group members cooperate with the information seeker and give predefined responses which allow to identify all information providers but one. When both approaches are combined, it might be possible to narrow the number of possible information seekers by analyzing the set of information providers – it is quite likely that the (hidden) information seeker is the user in the social network who connects the information providers (to reduce this risk, the information seeker should act as information provider herself to blur her activities). Furthermore, it must be ensured that the anonymous information seeker does not send a custom encryption key to each anonymously replying information provider, which allows to map the responses to the information providers (this could be done when all information providers receive the same message and mutually confirm to have the same encryption key for the response). Of course, the process does not protect the anonymity of information seeker or provider if they give hints about their identity in the message. Depending on the privacy concept, it could be possible to also include the following feature (*proof of social closeness*): when receiving an information need from an anonymous information seeker, the information provider has no information about the information seeker. To be able to distinguish requests from people within one's social network and requests from outside one's network, it could be useful to include evidence that the information seeker is socially close to the information provider. In a very basic form, this evidence could be implemented as a certain secret string the information provider shares with her social network (it could even be the information provider's public key used for encryption of the query) – however, it must be ensured that this feature is not used to exploit the privacy protection of the information seeker (by e.g. generating custom "proofs" for each social contact). However, this basic implementation can get circumvented by a social contact who shares the proof she received with others (so that they can use it to qualify as a valid social contact of the information provider).

8

INFORMATION SPACE

According to (Newby, 1996), an "(...) [information] space is a set of concepts and relations among them held by an information system". For the purpose of the social information retrieval approach proposed in this thesis, Newby's definition is too limiting. We therefore stick to the following definition: The private information space of a user consists of the user's information that has been explicated (i.e., that is available as materialized data) and the tacit information that is not digitally available but is only known to the user (and therefore stored in her brain). For the proposed system, it is not necessary to differentiate between *knowledge* and *information*. Research on knowledge management has several definitions for *knowledge* that range from "personalized information" to the process "of applying expertise" (Alavi and Leidner, 2001). For the social information retrieval approach discussed in the present thesis, the traditional IR paradigm is used: as a response to a previously asked query, a result set with information is presented to the information seeker. This information can be of any type (structured knowledge, problem descriptions, unstructured documents, etc.). The explicated information in the information space may include (but is not necessarily limited to) actively generated information (emails, SMS, reports, etc.), passively generated information (web browsing logs, received emails, transaction confirmations like order confirmations, locations), diverse multimedia information (e.g., text, image, audio, and speech), real world (personal papers, books, local librarians) and virtual information (web, digital libraries), and personal, organizational, public, and impersonal information. Based on the considered part of the information space, two distinct scenarios for social information retrieval are possible:

- Using the complete information space (requires the manual explication of information from the information provider)
- Using only already explicated information (would not necessarily require active involvement of the information provider, assuming that a working decision function with regard to privacy is in place)

In the following, the representation of information in the information space (Section 8.1) and the extraction of the information from its original system (Section 8.2) is explained.

8.1 REPRESENTING INFORMATION

8.1.1 *Structured Information*

When a data source reveals a clear, unambiguous assertion that can be either true or false, it is considered as a source for structured information. More strictly, this type of interaction can be seen as data retrieval (cf. Chapter 4). For those types of information sources, tools from the semantic web domain or traditional systems based on

predicate logic or databases are suitable to represent knowledge. To transfer information from the real world to the representation stored in the indexed information space, it is important to recognize the underlying structure of the data format to allow a proper mapping between the world and the representation. Typical examples for structured information types are location, transactional information (e.g., bought or viewed products, visited web sites, conducted web searches), people a person met, and meta information about other information items. In Section 8.2.1, extracting this information is discussed in further detail. For the sake of simplicity, the introduced structures are not related to existing semantic web technologies based on RDF or OWL (cf. Chapter 4) like SIOC[1], FoaF[2], or SKOS[3]. While it is easy to map the proposed concepts to the existing ontologies and structures, it could distract from the main ideas due to the additional overhead on the meta-level without providing any noteworthy benefits for this initial evaluation of limits and chances of social information retrieval. However, for a production-ready implementation of a social search system, this mapping is considered useful (e.g., to ease interoperability).

Storing structured information in the general record format introduced in Section 8.2.1 can be easily done using any relational database, like SQLite[4] or MySQL[5]. Both systems are able to deal with large amounts of data – however, for certain applications with heavy use of binary data, different, more specialized systems might be a better choice. Using internal database features like indexing will provide sufficient performance to deal with text-based records of structured data in the depicted social information retrieval scenario.

8.1.2 *Unstructured Information*

Connectors extract the information from the respective source system (e.g., email, web browser, etc.) to plain text. In the subsequent step, it is important to feed the data to the IR system which allows to integrate new information and query for information items. In the following paragraph, five different ways are discussed.

DATA PREPROCESSING For all subsequent steps this approach relies on the unigram model, where the probability for each word only depends on the word and not on the context (Manning et al., 2008, p. 240). For all texts, a set of commonly used preprocessing steps will be applied (Manning et al., 2008, p. 22f):

- Applying corpus-specific steps, e.g. normalization of character-encoding
- Removing punctuation, since it does not contain any information that would be useful in later steps
- Tokenization, i.e. breaking the text into its components (words). More complex representatives are language-dependent to be able to treat words according to custom language rules (e.g. disintegrate *aren't* or *isn't*).

1 `http://rdfs.org/sioc/spec/` (retrieved 2016-03-02)
2 `http://xmlns.com/foaf/spec/` (retrieved 2016-03-02)
3 `https://www.w3.org/TR/2009/REC-skos-reference-20090818/` (retrieved 2016-03-02)
4 `https://www.sqlite.org/` (retrieved 2016-01-17)
5 `https://www.mysql.com/` (retrieved 2016-01-17)

- Normalization steps ensure that all words are converted to upper or lower case, abbreviations are standardized (U.S.A. vs. USA) and that dates are represented in the same form.

While punctuation contains semantically rich information, it is not used in the following

TF-IDF Following the IR literature, a widely used approach to store and retrieve the data is the traditional TF-IDF approach (Section 2.3.2). Each vocabulary term constitutes a dimension in the vector space. The documents are represented as vectors with the magnitude of each of the vector's dimensions being calculated based on the well-known TF-IDF formula. Thus, the term frequency and the frequency of the term in the whole document collection are considered as a proxy for how important that word is for a document and how distinctive the word is within the whole document collection. Each query is translated to this vector space; using a similarity metric like cosine or Salton's measure (Section 2.3.2) it is possible to rank the documents in the information space for each query based on their similarity to the query. Often mentioned drawbacks of this approach are the bag-of-words assumption (the position of a word does not play a role in determining its importance for the overall document) and the vocabulary problem (only identical words cause a match – this is weakened by applying preprocessing steps like stemming). Due to its dissemination, it is considered as the baseline algorithm in the proposed scenario.

PROBABILISTIC LANGUAGE MODELS Approaches based on the probabilistic language model idea calculate how likely it is that a given query has been generated by a document's language model. The basic assumption is that a higher likelihood also translates into a higher relevance of the document for the given query. So, for a query q and a document d, one wants to estimate the probability that d is relevant for q, i.e. $P(d|q)$. Applying Bayes' formula, the probability translates to

$$P(d|q) = P(q|d) \cdot P(d) \cdot \frac{1}{P(q)}. \tag{60}$$

The first part of the formula ($P(q|d)$ can be interpreted as the probability that the query has been generated using the language model of document d, $P(d)$ is the document prior probability (how likely is it that this document is relevant), and $P(q)$ is the prior probability of the query.

Following the standard unigram model, $P(q|d)$ will be calculated based on the individual terms of q, i.e. $P(q|d) = \prod_i P(q_i|d)$ (cf. (Zhai and Lafferty, 2001)).

Without any further information, it seems reasonable to use term frequencies to estimate $P(q_i|d)$ with q_i being the atomic terms in q. The first part then becomes similar to $tf(t, d)$ in TF-IDF. There is, however, no IDF equivalent – unless smoothing is used to increase the probability for terms not explicitly present in a document (Zhai and Lafferty, 2001).

The second component of the formula, $P(d)$, defines the prior probability of the document (Miller et al., 1999). In our specific scenario, several sources to estimate the prior probability would make sense, since they rely on the individual social environment of the user:

- *Age of document*: Under specific circumstances, more recent information might be of higher importance (e.g. technical product information or discussions about trends). It can be useful to implement a penalty function to decrease the probability for older documents. Inspired by Li and Croft (Li and Croft, 2003), the age of a document can be used to decrease the likelihood of a document being relevant using an exponential distribution for $P(d) = \kappa e^{-\kappa(T_C - T_d)}$ where T_C is the date of the most current information item in the collection and T_d the creation date of the respective document (Efron and Golovchinsky, 2011; Li and Croft, 2003).

- *Access frequency of document (for specific query or in general)*: A document that has been requested often is certainly of better quality than a document that has not been retrieved at all. This factor could be modeled using a similar approach as above, giving an item a certain amount of "energy" when it gets retrieved, while the energy fades away when the document is not requested anymore. Potential candidates for mathematical representations can e.g. be a linear ($P(d) = 1 - \frac{\Delta t_d}{\Delta t_{max}}$) or an exponential function ($P(d) = 1 - \exp(-(\Delta t_{max} - \Delta t_d))$) with Δt_d being the elapsed time since the last access of document d and Δt_{max} being the maximum elapsed time since the last retrieval in the collection (considering only the documents that got retrieved at all).

- *Privacy considerations*: How often has this document been shared already with others? If a user did actively not share the document that has been recommended for sharing, this behavior might indicate that the document is either not of the desired quality or contains information that the information provider is not feeling comfortable to share it with others. An approach to quantify this factor in terms of probabilities is

$$P(d) = \frac{|\text{people}_{\text{shared}}|}{|\text{people}_{\text{recommended\_sharing}} \cup \text{people}_{\text{shared}}|},$$

  where $\text{people}_{\text{shared}}$ is the set of people the user shared the information item with and $\text{people}_{\text{recommended\_sharing}}$ is the set where it was recommended to share the information item. This ratio gives an indication how likely it is that the item is shared by the user when the algorithm decides that the content would help to satisfy a certain query. One has to admit that the ratio does not only reflect the user's privacy/sharing preference for the respective item but also implicitly measures the system's algorithm to find suitable information that would fit a query (imagine a situation where a user decides not to share a certain item because it does not fit from a content perspective). This could be mitigated by designing the user interface in a way where the user indicates why a proposed item is not shared with the information seeker. Following this approach, it is possible to clearly distinguish both cases and to adjust the formula as

$$P(d) = \frac{|\text{people}_{\text{shared}}|}{|\text{people}_{\text{not\_shared\_due\_to\_privacy}} \cup \text{people}_{\text{shared}}|}$$

  with $\text{people}_{\text{not\_shared\_due\_to\_privacy}}$ reflecting the set of people with whom the information item has not been shared for privacy reasons.

- *Co-requests*: How often has this document been retrieved along with the other documents in the result set? Assuming that D is the set of documents in the information space, the power set 2^D is the set of all potential result sets that can get returned to an information seeker. Furthermore, let $c(r) : 2^D \mapsto \mathbb{N}_0$ define a function that assigns each potential result set the number of times it has been returned as a response to a query. The probability that a document d_1 is relevant when documents $\{d_2, ..., d_n\} \in D_{relevant}$ have already been identified as relevant can be quantified based on previous result sets, namely as

$$\left(\sum_{\{d_1 \cup D_{relevant}\} \subseteq t \in 2^D} c(t) - \sum_{D_{relevant} \subseteq u \in 2^D, d_1 \notin u} c(u) \right) \cdot \frac{1}{\sum_{j \in 2^D} c(j)}.$$

  Following this approach, the appearances of result sets including $D_{relevant}$ are subtracted from the the number of times where the result set also contained d_1. The difference is normalized with $\sum_{j \in 2^D} c(j)$.

To summarize the previous paragraph, a document's prior probability can be estimated based on the mentioned criteria age, popularity, active involvement, and co-requests using a linear combination

$$P(d) = \sum_i \lambda_i \cdot p_i \tag{61}$$

with p_i being the probability for each of the influencing factors above and an additional default probability, $\frac{1}{|D|}$, the prior probability if all other λ_i are set to 0. With λ, it is possible to adjust the impact of the single components.

The third component of the formula is the prior probability of the query. Since it is independent from the document corpus, it is treated as a constant and therefore not discussed further.

LATENT DIRICHLET ALLOCATION Latent Dirichlet Allocation has been introduced in Section 5.4.1. It allows to create a vector space based on the latent topics that have been identified in a large document collection, e.g. the user's information space. Since each document or query can be expressed as a vector, classical metrics like cosine or Jensen-Shannon divergence (Section 2.3.2, Section 5.4.2) can be used to calculate the distance between two vectors. The topic space can be built based on the user's document collection directly or using a background collection, like e.g. Wikipedia. Imagine a scenario where a user has a large collection of computer science documents. Following the former approach (building the topic structure using the document collection directly), the topic structure represents the document collection very closely, i.e. it is quite likely that computer science would be organized in many subcategories, whereas it would only be one category for someone who does not have that many documents about computer science. However, topics that are not covered in the document collection are not considered in the topic space. This could be harmful when querying the information space for content that is not represented sufficiently. Following the latter approach (building the topic structure using a background collection) would result in a topic structure that is much more balanced and "universal" since it

is not tightly coupled to the user's information space (in fact, every user could use the same topic structure). The user's documents are put into perspective and context, which helps to identify topics that are quite rare in the information space, at the cost of lower coupling to the user's document collection. In the specific use case of this concept, however, one could argue that the common social context possibly would imply sufficient content overlap among the users to ensure that a more granular and individual representation of topic areas brings more benefits than the broader categorization of topics that would result from using an additional source to create the topic space. In addition, previous research (Cimiano et al., 2009) shows that using an additional corpus to calculate the topic space has been proven to be worse than an explicit approach like ESA (Section 5.5). A potential compromise of both extreme cases introduced above could be a hybrid system that uses a model based on a more general collection like Wikipedia if the model directly built from the user's collection does not provide sufficiently high θ values for the query (i.e., the linkage between the topics and the query is low).

PROBABILISTIC LANGUAGE MODELS WITH LDA SMOOTHING One of the problems immanent to all approaches that are based on a vocabulary concept is that two different words are orthogonal to each other, no matter how close the two words might be from a semantic perspective: if two words share the same meaning, but are represented by different terms, they still are considered different. Some approaches propose LDA to "smooth" the retrieval metrics, by combining elements from the probabilistic models (term similarities) and concepts from LDA (similarities in the topic space, identified using patterns of term co-occurrence). An exemplary approach is explained in (Zhai and Lafferty, 2001).

EXPLICIT SEMANTIC ANALYSIS A relatively new approach to IR is to represent documents and queries as vectors in a concept space defined by Wikipedia articles (Explicit Semantic Analysis, ESA, cf. Section 5.5). An advantage is that all users share a similar structure for their information space, so the representation of queries is the same for all information spaces and no transformation steps are needed to exchange vectors among users. In addition, loss of information during the matching step can be prevented, since the approach allows to represent all queries in the concept space, even if there is no matching document in the user's information space. Furthermore, all dimensions of the concept space are easily understandable (since they represent well-known concepts, expressed as articles on Wikipedia).

A potential disadvantage is that the structure of topics is not fitted to the user's collection of information items, which might result in a less fine-tuned category clustering. However, since the underlying corpus (i.e., Wikipedia) can be considered broad enough and has been proven to outperform other approaches (Cimiano et al., 2009), the disadvantage needs to be considered in context. An additional disadvantage is the dependency on an external corpus, which has a huge impact on the overall quality of the system – especially over time when new concepts occur (e.g. due to technical progress or environmental changes).

Given that nearly all approaches have their individual benefits and shortcomings, a variety of them will be compared and discussed in the evaluation section (Part III).

8.2 EXTRACTING INFORMATION

Information spaces as specified above are flexible to deal with several information types. This section will discuss common examples for specific types of information in detail and show exemplary ways to collect this information. Due to their different properties, structured and unstructured information are treated separately. It is important to note that the proposed social search concept can get enhanced by including other information domains (like, e.g., processed audio or video) as well. For the sake of explaining, developing, and evaluating the basic concept, this work focuses on the examples below since those data types are relatively easy to model and still allow to run a set of useful example scenarios for the social search concept.

8.2.1 *Structured Information*

Typical examples for structured information types are location, transactional information (e.g., bought or viewed products, visited web sites, conducted web searches), people a person met, and meta information about other information items. In the following paragraphs, the collection process of these information items is discussed in detail.

A generic data record *r* for structured information can be written as 5-tuple *r=(id, (date,time), type, data, meta)* where *id* is an identifier (e.g. an auto-incrementing number), *(date,time)* defines temporal coordinates of the event, *type* indicates the record's type (incl. a link to a full type specification), *data* represents the actual payload, and *meta* contains additional metadata. The content of all fields is defined in the specification referenced in the *type* field for each record type (e.g., providing a format description for *data* and *meta* fields). Assuming that */types/location* points to a specification for spatial coordinates, a trivial example for a data record which represents a location could be *(1, (2015/09/29,17.35 CET), /types/location, (48.257175, 12.522306), "home")*.

LOCATION Obtaining location information of a user with GPS sensors in mobile devices is not a major challenge anymore. To convert these coordinates to semantically meaningful places (e.g. restaurants, shopping malls, garages, etc.) to fill the *meta* field introduced above, various approaches have been proposed (e.g. (Uzun et al., 2013; Liu et al., 2006; Cao et al., 2010; Nurmi and Bhattacharya, 2008)). For the location connector, it therefore is valid to assume that it is possible to implement a software program that captures the user's current physical location, translates that location to a semantic representation, following a flat but standardized ontology, and saves the record including a time stamp.

TRANSACTIONAL INFORMATION Examples for structured transactional data include visited websites or bought products. Both cases can be easily modeled using the generic data record format above: a visited website, e.g. *http://www.cs.tum.edu/*, can be simply represented as *(1, (2015/09/29,19.03 CET), /types/url, http://www.cs.tum.edu/, "")* while a bought product needs to relate to a common structure to identify the product. To facilitate downstream interpretation of the data, it is important to avoid

ambiguity in the product definition wherever possible. A simple string matching of product names is not sufficient, since vendors sometimes name their items differently (e.g. when bundling them). Furthermore, it is not possible to locate the item in a specific structure, e.g. it is not obvious that a product labeled as "Nikon D610" fulfills the same purpose as a product named "Canon 6D". A semantic representation of products and their relations (e.g. via product categories) would greatly improve the usability of the social information retrieval application (see example use cases in Chapter 9). A widely used system to classify goods is the Global Product Classification (GPC) system which relies on Global Trade Item Numbers (GTIN) and has been explained in Chapter 4. GPC defines a category tree for goods and allows to locate items within this category tree. A much more pragmatic approach would be to use Amazon's system of Standard Identification Numbers (ASIN) and categories: as explained in Chapter 4, each product that is listed on Amazon has a unique ID, referred to as *Amazon Standard Identification Number* (ASIN). Using this number, one can retrieve the product description by downloading http://www.amazon.com/dp/*ASIN*. Furthermore, Amazon uses a flat category system, where the name of a product's category is printed in the HTML title-tag of the product's web page in the Amazon online store. A disadvantage of this approach is that the category names are translated to the respective local language, so *http://www.amazon.**de**/dp/B0000 5N5PF* lists *"Spielzeug"* as category, while *http://www.amazon.**com**/dp/B0000 5N5PF* assigns the item to the *"Toys & Games"* category. An easy workaround could be to create translation tables. Nevertheless, the Amazon system is not meant to be used that way: while it is simple and powerful enough to be used as a proof of concept in this thesis, a real-world solution should be built upon something more mature that is either an open standard (like Wikipedia, cf. Section 5.5) or the aforementioned GPC system.

SOCIAL CONTACTS The set of people who are part of the user's social situation can be seen as an additional form of transactional information. In (Groh et al., 2011b), the authors proposed a framework to combine multiple evidences to estimate the current social situation a person might be engaged in. Therefore, a social situation is defined as 4-tuple $S = (P, T, X, K)$ with P being the participants of the situation (i.e. the other people involved in the social situation), T representing a temporal reference, X defining a spatial reference and K providing information about the social situation's semantic. Using the generic record format explained above, including this kind of information would only require a sufficient clear definition of the format and a working implementation of the sensors.

Previous studies suggest that obtaining individual information is possible from a technical perspective, so this work will not discuss these parts in detail. In (Murakami et al., 2012), a data structure of time, keywords, and URI sets is used to help users to memorize certain days in their life, based on information from their web searches, twitter, emails, calendars, or book purchases. (Hangal et al., 2011; Hangal, 2012) rely on email archives for reminiscence.

8.2.2 *Unstructured Information*

The second type of data indexed on the user agent is unstructured data. Unlike the various forms of structured data discussed above, this type of data does not follow a clear logical structure. It is much more difficult to make sense of – e.g., while inference is possible in predicate logic using structured dataset, unstructured information includes much more uncertainty in the reasoning process. Examples for unstructured information include the consumed web information (i.e., the content of pages that have been browsed) or other information that has been read (e.g., locally stored PDF documents, content of email conversations, etc.). Figure 4 shows a small set of possible examples for unstructured data sources: web history, local files, and content of communication done via Twitter, Facebook, or Email have in common that it is difficult to represent their content in a way that would fit to logic based frameworks. Therefore, using topic modeling approaches like LDA (Section 5.4.1) or ESA (Section 5.5) might help to reveal underlying topics and constitute a more flexible way of making use of the available information. Extracting the information from their original systems is done using connectors. In the remainder of this chapter, some examples are explained in detail. The list is not considered to be exhaustive, especially because new sensor technology like Google Glass could increase the amount of available data tremendously: with more information being available digitally, everything the user sees could be directly processed and considered as part of the explicated information space. Transforming the input of optical or acoustic sensors to a reasonable format is not the topic of this thesis, therefore this approach relies on the text representation of knowledge. In the end, different options for storing the data and building an index structure are discussed (LDA, LDA+TF-IDF, ESA).

As a general framework, each of the approaches conducts the following steps:

1. Get access to the source of the documents
2. Extract the text of the source
3. Index the results
4. Maintain and update the index to reflect changes

VISITED WEBSITES The significant part of consumed knowledge is received via online media[6]. Therefore, using the information read online by a user to estimate the user's information space seems a plausible approach, since it covers a large part of the user's individual knowledge gain.

Obtaining a list of visited websites and reading their content can be achieved referring to the browser's history and/or cache. Conventional browsers maintain a list of visited URLs and often even store the HTML code of the displayed websites. Developing a tool to either access the browsers' history to re-download the visited URLs or to read the browser's cache directly and evaluate the content without additional re-download is a task that can be done for the majority of available browsers. For Chrome and Firefox, a proof of concept has been developed to demonstrate the idea (cf. Section A.2.2 as part of Experiment 7, cf. Chapter 16). To remove the HTML

6 http://www.statista.com/statistics/422572/europe-daily-media-usage/ (retrieved 2016-01-11)

markup, well-established libraries like the Natural Language Toolkit (NLTK)[7] offer proven methods to extract the plain text.

READ DOCUMENTS Documents stored on the user's computer build the next corner stone to assemble a user's information space. Unlike websites, files might not have been published, causing the available information for search to increase in comparison to normal search engines. Aside from rare exceptions, the documents stored on a user's computer represent a part of the user's knowledge. Obtaining the documents is not challenging, since all documents in question are stored on the user's computer. Section A.2.2 describes a proof of concept (reading PDF files that got downloaded from the web and stored locally on the user's computer).

COMMUNICATION Previous studies suggest that personal email archives can contain valuable information for the search process (Nagpal et al., 2012); therefore, communication logs and content should be included. Extracting content from email depends on the email provider, but can be trivially done using a custom-made IMAP or POP3 client to download the messages in text format. Twitter allows to download a user's tweets using an API[8], while Facebook also offers a backup functionality to retrieve all data stored in one's profile (cf. Section A.2.2).

8.3 PRIVACY

To protect the information provider's privacy, a respective privacy function is assumed to exist. This function decides for each triple (information provider, information seeker, information item) whether the information provider shares it with the information seeker or not. Formally, it is a function s which returns true or false (i.e., to share or not to share), depending on information provider, information seeker, information item, and query:

$$s : \text{User}_{\text{IP}} \times \text{User}_{\text{IS}} \times \text{Information Item} \times \text{Query} \mapsto \{\text{true,false}\}. \tag{62}$$

The function s is supposed to exist. In the absence of such a function, the information provider decides manually whether an information item can get shared with an information seeker.

7 http://www.nltk.org/ (retrieved 2016-01-13)

8 https://dev.twitter.com/overview/api/twitter-libraries (retrieved 2015-10-21)

9

CORE SERVICES AND USE CASES

The following section provides an overview of potential use cases for an implementation of the proposed social search concept. The scenarios are inspired to a modest degree by the results of Experiment 2 (cf. Chapter 11) and (Oeldorf-Hirsch et al., 2014). Social search appears to be beneficial for information needs which

- precede the information seeker's decisions (e.g. decision to buy a certain product, to consult a certain doctor/lawyer),
- are highly individual and contextually dependent (and therefore often have no general answer),
- are of a rather complicated nature (i.e., are too complex for the user to be specified completely as query in a traditional search engine), and/or
- have an important relation to social context.

9.1 STATISTICS ON FRIENDS' ACTIVITIES

Relying on social concepts like homophily, influence, and cofounding introduced in Section 2.4.2, the following example shows how decisions of individual users can be accelerated and – following the social correlation theory (Tang et al., 2014) – partially supported.

When assessing the different options in preparation of an important buying decision, a user has several sources of information: publicly available product descriptions, reviews and articles from magazines or websites, reports about other users' experience (e.g. published on the web), and recommendations from experts (e.g. store clerks) or other people. The whole process of collecting the information and inspecting it thoroughly is time-consuming and requires much effort. In many cases, people therefore use shortcuts to come to conclusions, e.g. by restricting the search process to a very limited number of information sources. Having the possibility to get a clustered overview of typical products (matching the requirements) bought by people within one's social environment could improve the decision making in several ways, depending on the result:

- Increase decision making pace by selecting one of the dominant options in the circle of friends (e.g., if the information seeker needs to decide fast and/or the preference of the information provider is quite clear).
- Increase the amount of information that is available to come to a decision. The information seeker can take note of the friends' decisions and can consider them in her own decision. The additional information could be of higher interest for the information seeker, due to the additional trust that is caused by social closeness (a sales person is normally incentivized by increasing sales volume,

while a socially close friend might have a higher interest in helping the information seeker). In addition, serendipitous effects might occur with a higher probability due to social closeness and homophily (cf. Section 2.4.2): the fact that people within the information seeker's social network decide in a specific way (previously not perceived as an option) might influence her own decision in a previously unforeseen way.

For privacy reasons, decisions of information providers could only be presented in aggregated form to make drawing conclusions about individual information providers more difficult.

The information seeker defines a query q for a product and sends the request to all direct social contacts DIP based on her social network. It is assumed that the information seeker has access to an explicit representation of her social network (cf. Section 7.1). It is possible to modify the set of considered contacts DIP at query time, and to include more distant users, e.g. second-degree contacts. The query $q =< type, product >$ is linked to an ontology, representing the type of query (e.g., "buying support") and the product in question. Furthermore, the proposed concept relies on a product ontology, different options have been presented in Section 8.2.1 and Chapter 4.

The designated information providers (members of set DIP) use the search functionality of their information space (Chapter 8) and identify potentially relevant information items. Following an (optional) privacy-aware collection protocol summarized in Section 7.2.2, each information provider sends her reply (directly or indirectly) to the information seeker, who is able to aggregate the information.

Depending on the responses, the information seeker can aggregate the data according to several dimensions, e.g. product, time, or price. It might be valuable to see whether buying decisions changed over time, whether there are clusters of items which got bought, or items that have been considered by a user but not chosen in the final step. Thinking ahead, additional meta-information (usage patterns, overall satisfaction) could be added to the data type as well and increase the benefits of the service.

9.2 EXPERTISE IDENTIFICATION

It is possible that members of one's social network have expertise in areas one would not think of. Only due to random serendipitous situations tacit knowledge eventually becomes apparent and could either foster mutual exchange or could possibly be used for problem solving.

A social search system could help finding possibly rare expertise within social reach – in some cases, it might already be helpful to consult someone with at least a decent amount of expertise (but who is available and trustworthy) than identifying the ultimate expert of a knowledge domain (who will most likely not reply to the request, (Kukla et al., 2012)).

Depending on the specific form of creating the knowledge profile introduced in Section 7.1.2, the profile is built based on the other user's consumed information and therefore could reflect her areas of expertise quite well. Identifying areas of expertise based on web browsing can only be seen as a first approximation – but the fact

that someone stores documents about or searches the web for a certain topic clearly demonstrates a certain degree of interest, which might indicate expertise.

9.3 EXPERTISE GAP IDENTIFICATION

Exchanging individual knowledge profiles within a social network allows for finding areas of uncovered expertise. By calculating a knowledge profile vector for the group (using simple vector addition) it is possible to easily reflect the existing knowledge within the group. Aggregating the group knowledge profile vector to a higher level of abstraction (e.g., for ESA, using the underlying semantic structure of the vector space's dimensions to condense the vector to central concepts only), it could become possible to interpret the vector and compare it with a target state derived from the group's mission. This process could reveal areas of improvement and thereby generate opportunities for the new member. Potential use cases could be working teams, such as research groups at universities or companies.

9.4 BOOKKEEPING SYSTEM / MARKET APPROACH

According to (Dunbar, 1992; Dunbar, 1993), homo sapiens has a cognitive limit for the number of people with whom she can maintain stable social relationships. The figure, that became famous as "Dunbar's number", is based on a co-evolution of neocortex size and group size of primate populations. For humans, the predicted group size is around 150. Today, online social network platforms like Facebook help to "preserve" social relationships and thereby overcome our mental limitations. A study[1] reveals that on average, Facebook users have much more friends than Dunbar's number would suggest (total average: 350[2]). A social search market model could keep track of mutually exchanged favors and preserve "invested social capital". The market model would keep track of exchanged information and keep a record of perceived deposits and debts. The rating would be done individually by each user and would not be shared. It is intended to provide individual support to memorize past exchanges and should not start a negotiation about the value of certain actions. A system could use the market model introduced in Section 7.1.3 to provide evidence-based recommendations for the user's decision process.

1 http://www.statista.com/statistics/232499/americans-who-use-social-networking-sites-several-times-per-day/ (retrieved 2016-01-14)

2 Average number of friends split by age groups: 12-17: 521, 18-24: 649, 25-34: 360, 35-44: 277, 45-54: 220, 55-64: 129, 65+: 102

Part III

EMPIRICAL STUDIES

The following chapters describe the empirical experiments that have been conducted to evaluate the open variables and their potential relationships of the concept introduced in Part II. The experiments are designed to cover the complete user journey of a social information retrieval session:

- Experiment 1 (Chapter 10) suggests that information items in friends' information spaces constitute valuable objects for information retrieval. A large dataset collected from Twitter and Facebook is used to show that information published by socially close people is of higher interest for a user than information published by someone else.
- Experiment 2 (Chapter 11) indicates that users feel much more comfortable to ask others for information when they can explicitly define the recipients of their query. The results are obtained conducting a survey with 112 participants.
- Experiment 3 (Chapter 12) investigates whether certain content areas are better suited for social information retrieval than others. A manually collected corpus of websites from various content areas is rated according to several dimensions by members of a crowdsourcing platform to identify patterns which might explain why information needs with certain characteristics are more "social" than others.
- Experiment 4 (Chapter 13) evaluates the performance of various mechanisms proposed in Part II to identify a set of potential information providers. The experiment focuses on expertise (and omits the concept of the social capital market introduced in Section 7.1.3) and compares TF, TF-IDF, topic models (LDA), and ESA-based mechanisms using the Cranfield collection, a set of 226 predefined queries and related relevance judgments for 1,400 scientific abstracts.
- Experiment 5 (Chapter 14) and 6 (Chapter 15) compare the performance of several ways to organize the information provider's information space. Using a publicly available dataset obtained from the Stackexchange communities, TF-IDF, various LDA-based approaches, and ESA are used to identify relevant answers for questions.

- Experiment 7 (Chapter 16) is an experiment with > 100 students over a time period of several weeks. While the other experiments cover smaller, more isolated parts of the social information retrieval concept, this experiment was designed to provide a holistic perspective on the two technical scenarios (the manual approach, with manual routing of queries and solely manual answers given by the information provider and the fully automated approach, where routing is done by the system and the answer is based on the information provider's digital information space) and a specific use case (Social Product Search). It therefore is structured in Experiment 7a (Manual Approach, cf. Section 16.5.2), Experiment 7b (Automatic Mode, cf. Section 16.5.3), and Experiment 7c (Social Product Search, cf. Section 16.5.4).

Some contents of this part have already been presented at conferences and published in the respective conference proceedings: in (Fuchs et al., 2015) the dataset, evaluation approach, and results for Experiment 1 are explained, in (Fuchs and Groh, 2015b) the study setup and results from Experiment 2 are described, in (Fuchs et al., 2016a) the results of Experiment 3 are presented, in (Fuchs and Groh, 2016b) Experiment 4 is discussed, in (Fuchs et al., 2016b) Experiment 5 and 6 are explained, and in (Fuchs and Groh, 2015a) the setup for Experiment 7 is discussed in detail. The results of Experiment 7 have also been discussed in (Fuchs and Groh, 2016a).

10

EXPERIMENT 1: ELIGIBILITY OF FRIENDS' INFORMATION SPACES FOR INFORMATION RETRIEVAL

The datasets used in this experiment have been collected as part of Gregor Semmler's Bachelor's Thesis (Semmler, 2013) and the Master's Theses written by Florian Hartl (Hartl, 2013) and Benjamin Koster (Koster, 2013). All three theses have been supervised by Jan Hauffa and Georg Groh at the Chair for Applied Informatics – Cooperative Systems at Technische Universität München. Parts of this section appear also in (Fuchs et al., 2015).

10.1 SYNOPSIS

When users try to satisfy information needs, relevance is traditionally defined by metrics based on term or concept distance (Manning et al., 2008), link structure of the investigated information collection (Brin and Page, 1999; Kleinberg, 1999), or selected results in previously conducted search sessions (by the same or other users) (Micarelli et al., 2007). Leveraging the information within one's own social network to enrich search results is currently discussed in the specific forms of Social Media Question Asking (SMQA) and Social Search (Oeldorf-Hirsch et al., 2014; McDonnell and Shiri, 2011). Analyzing two large datasets crawled from Twitter (360,000 user profiles, 223 million tweets) and Facebook (25,737 user profiles, 4.6 million posts from 936,992 users), our findings suggest that content created by people who are socially close is of higher individual relevance than content created by others. Furthermore, our results indicate that the willingness to help satisfying information needs is higher for users within one's social network.

10.2 MOTIVATION

Several approaches have been discussed to enhance the search and recommendation process with social aspects (Oeldorf-Hirsch et al., 2014; Lampe et al., 2014). Traditionally, search results are calculated based on several relevance metrics based on search term distance, link structure, historic information (e.g. for personalization and former relevance judgments) and ontology-based approaches for conceptualization. Taking "social" in Social Search (cf. Section 2.5) seriously, relevance must be regarded in a much broader sense, allowing social influence to impact individual relevance judgments like in marketing (Burnkrant and Cousineau, 1975) or in the choice of apps on a mobile phone (Aharony et al., 2011). Information items might at first only be relevant because friends are interested in them but also could provide more serendipitous results caused by the social relevance of these items in the social groups of a user. Thus, they belong to the sphere of the wider unconscious information needs (cf. (Groh et al., 2013), Section 2.2, and Section 2.3.5). A more "social" search engine could allow users to query privacy restricted, non-public information spaces of their friends directly (as proposed in Part II), leading to results with a special social fla-

vor of relevance (due to highly individual knowledge about the information seeker) and a broader information space (due to access to otherwise restricted information). Based on a corpus crawled from Twitter we investigate retweet behavior, considering retweeting a message as a positive relevance judgment. Our findings indicate that users assess tweets of people they directly interacted with as more relevant than messages sent from others. Furthermore, our analysis suggests that people we interacted with before tend to react to our questions faster than others. Analyzing messages posted on Facebook, we can show that replies to a question are liked more by the information seeker (and are therefore possibly considered to be more helpful) when the reply is posted from within one's own social network.

10.3 RESEARCH QUESTIONS

The experiment focuses on the following research questions:

I. Are relevance judgments on content correlated to the strength of the social relationship between author and recipient of the content?
II. Does social closeness influence the willingness to react to questions in a social media question asking scenario?

A positive relationship in RQ I suggests that search applications could benefit from integrating the content of socially close friends. Affirming RQ II indicates that a distributed social search approach should rely on querying socially close people.

10.4 DATASET

Twitter is one of the major online social networking services with more than 200 million active users by the time the dataset was collected[1]. Users have the possibility to select a tweet and resend it to their own followers (i.e., *retweet* it). In addition, users can send direct (but public) messages to other users using Twitter's @-operator at the beginning of a tweet. The dataset consists of a large sub-graph of Twitter on a per-user-basis by means of breadth-first search (Granovetter, 1976), containing publicly available tweets dated between January 1 and July 26, 2012.

Facebook is one of the world's largest online social networks. Among numerous other things, users can establish friendship edges and post (and reply to) public messages. Users also have the ability to *like* content objects. The Facebook dataset has been retrieved using a crawling procedure based on Metropolis-Hastings Random Walk (cf. (Semmler, 2013; Gjoka et al., 2009)).

Tweets/posts ending with a question mark were regarded as questions. Previous research (Teevan et al., 2011) analyzing response quality and quantity in SMQA showed that phrasing a question as a single sentence with a question mark improves response quality and quantity. Validity checks on subsets (100 posts/tweets) revealed recall / precision values of 83%/55% (Facebook) and 25%/66% (Twitter)[2] in identifying questions. A question was considered as a "relevant" question only in case one could expect a real answer (e.g., no rhetorical questions). We are not in a position to reliably check to which degree a reply is a valid response to a question – even if the reply

1 `https://twitter.com/twitter/status/281051652235087872` (retrieved 2016-01-17)

2 The lower recall value on Twitter is caused by the heavy usage of hashtags following the question mark

is a counter question it could provide information that is considered as helpful by the author of the question. A small sanity check of 100 randomly chosen questions from the Facebook dataset and their respective answers confirmed that the answers in general fit the questions. The dataset only consists of publicly available data, accessible by any internet user. The data is not published and is only used for legitimate scientific research. The published information derived from the data does not disclose any details about the crawled profiles.

10.5 EVALUATION METHODS

10.5.1 *RQ I: Correlation of Relevance Judgments and Depth of Social Relationships*

TWITTER We consider the act of retweeting a tweet as a relevance judgment in favor of the original tweet by the retweeting user and assume that two users are "socially connected" if they have exchanged at least one directed message using Twitter's @-operator (regardless of the direction of the message). *Following* a user is not considered as a form of social connection, since it is one-sided and often motivated by the posted content (and not the respective "real" person). To simplify further analysis, we stick to the following notation:

- M_b^a is the number of directed posts (using Twitter's @-operator) sent from user a to user b,
- RT_b^a is the number of tweets originally sent from user b and retweeted by user a,
- RT^a is the total number of all retweets sent from user a, i.e. $RT^a = \sum_{x \in U} RT_x^a$ with U being the set of all users, and
- TW^a is the number of tweets sent by user a.

RT_b^a and TW^a are also defined for a set of users U, i.e. $RT_U^a = \sum_{u \in U} RT_u^a$ and $TW^U = \sum_{u \in U} TW^u$. The set of retweets posted by a user u may contain tweets which were originally posted by (1) users who are followed by u, (2) users who exchanged at least one direct message with u and (3) users who do not belong to either of the previous two groups. Therefore, we define the following sets of users:

- Friends(u) are users who are followed by u,
- SocConn(u) are users who exchanged at least one direct message with u,
- Other(u) are users who do not belong to either group, i.e. $\overline{\text{Friends}(u) \cap \text{SocConn}(u)}$

Due to Twitter's API limitations, it is not possible to reconstruct the retweet graph: if a user a originally tweets a tweet t, a different user b retweets t as t' and a third user c retweets t' as t'', the tweet t'' is only marked as a retweet of t (but not of t'). Therefore, it is possible that users appear to retweet tweets from strangers (i.e. users not connected via follower edges or direct messages).

To quantify retweet ratios, we use the function $R_1^u(x)$ defined as

$$R_1^u(x) := \frac{RT_x^u}{RT^u_{\text{SocConn}(u) \cup \text{Friends}(u)}} \tag{63}$$

which represents the ratio between tweets originally from users of group x which were retweeted by u and tweets originally from users within u's social contacts and people u follows which also was retweeted by u. A broader indicator, $R_2^u(x)$, also covering tweets from people where no connection exists, is defined as

$$R_2^u(x) := \frac{RT_x^u}{RT^u_{SocConn(u) \cup Friends(u) \cup Other(u)}} \tag{64}$$

Assuming that retweets are distributed equally, it is useful to compare $R_1^u(x)$ and $R_2^u(x)$ with $T_1^u(x)$ and $T_2^u(x)$ reflecting the contribution of the respective group to the overall tweet corpus. $T_1^u(x)$ is defined as

$$T_1^u(x) := \frac{TW^x}{TW^{SocConn(u) \cup Friends(u)}} \tag{65}$$

It represents the ratio of tweets posted by the respective group within the collected corpus, excluding users who are not followed or socially linked. This definition can be extended in analogy to R_2^u and is defined as

$$T_2^u(x) := \frac{TW^x}{TW^{SocConn(u) \cup Friends(u) \cup Other(u)}} \tag{66}$$

While $R_1^u(x)$ and $R_2^u(x)$ represent the contribution of a particular group (socially connected people; friends, i.e. people one follows; strangers) to the set of tweets retweeted by user u, $T_1^u(x)$ and $T_2^u(x)$ represent the proportion of all tweets posted by the respective group within the corpus. If one of the groups is overrepresented within the set of retweets (in comparison with the group's proportion in the full corpus) it could suggest that this group has more relevant content for a user u than other groups. To quantify this overrepresentation, the average ratios $1/|U| \cdot \sum_{u \in U} R_k^u / T_k^u$ are used for $k \in \{1, 2\}$ with U being the set of all available users.

FACEBOOK For our analysis, we rely on the existing friendship network within Facebook and interpret the *likes* of a user as a relevance judgment. We (1) identify the set of questions and (2) analyze the responses for identified questions, i.e. check whether the response has been posted by a friend of the question asker and check whether the question asker liked the response. A higher *like*-ratio for responses written by friends of the question asker than for other responses could suggest that friends are able to provide more valuable information than strangers. Using an explicit relevance judgment of the asker is in line with previous research on community Q&A (e.g. (Shah and Pomerantz, 2010)), where the answer explicitly chosen by the asker is considered best.

10.5.2 *RQ II: Relation of Willingness to Help and Social Closeness*

TWITTER To assess the willingness to help other users, we analyze tweets containing a question indicated by ending with a question mark. For each of these tweets, we identify the responses and calculate response time and number of messages sent

Group	R_1	T_1	R_2	T_2
Friends ∩ SocConn	0.35	0.07	0.27	0.07
SocConn	0.41	0.15	0.31	0.13
Friends	0.94	0.93	0.71	0.89
SocConn ∩ $\overline{\text{Friends}}$	0.06	0.07	0.04	0.06
Friends ∩ $\overline{\text{SocConn}}$	0.59	0.85	0.44	0.82
Others	n/a	n/a	0.25	0.05

Table 5: R_1, T_1, R_2, and T_2, averaged over all users (Experiment 1)

between the user asking the question and the user providing the answer. A negative correlation between response time and the number of exchanged directed messages would suggest that the closer two users are (i.e., the more direct messages they exchanged), the faster they reply to each others' questions. To improve the likelihood of the selected tweets actually forming a relevant question/answer pair, we only considered pairs of tweets posted within a time span of less than three weeks.

10.6 RESULTS

10.6.1 *RQ I: Correlation of Relevance Judgments and Social Connections*

TWITTER The average set of retweets of a user consists of tweets from users in the sets Friends ∩ $\overline{\text{SocConn}}$ (44.4%), Friends ∩ SocConn (26.7%), Others (24.7%) and SocConn ∩ $\overline{\text{Friends}}$ (4.1%). Table 5 shows the average results for R_1, R_2, T_1 and T_2 for the respective groups. Figure 5 depicts the ratio $\frac{R_1}{T_1}$ which can get interpreted as the degree of overrepresentation of a specific group within the set of retweets. It is noticeable that users retweet tweets from users who they follow and exchange direct messages (SocConn ∩ Friends) with much more often than their contribution to the overall amount of tweets would suggest. Furthermore, users one exchanged messages with, but did not follow (group SocConn ∩ $\overline{\text{Friend}}$), were retweeted (relatively) more often than users one only followed (group Friend ∩ $\overline{\text{SocConn}}$; 0.88 vs. 0.69).

FACEBOOK Out of 87,268 replies, 73,941 replies came from friends (thereof 11,144 were liked by the question asker) and 13,327 were given from other users (thereof, 1,692 were liked by the question asker). On average, 15.1% of the answers are liked by the question asker if the author of the response is marked as a *friend* on Facebook – if this is not the case, the question asker likes only 12.7% of the replies. Fitting a linear regression model revealed a significant, but very weak positive correlation ($\text{ASKER\_LIKES} = 0.02 \cdot \text{IS\_FRIEND} + 0.13$, with $p = 0.00$ but a very low R^2 score of 0.0006).

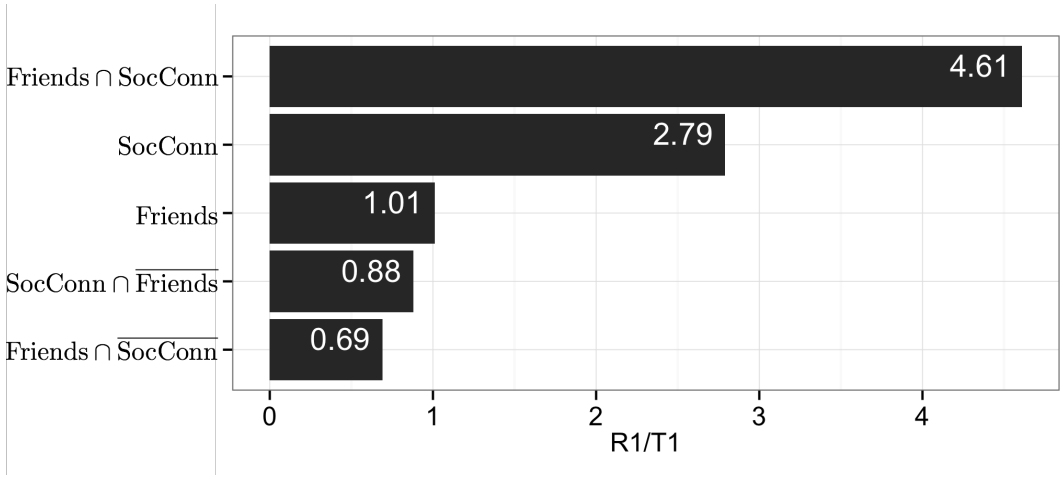


Figure 5: Degree of overrepresentation of groups within the retweet set (R_1/T_1) – users in the intersection of social connection and friendship (i.e. users who are being followed and exchanged at least one directed message) are retweeted 4.6 times more often on average than their overall contribution would suggest

10.6.2 *RQ II: Relation of Willingness to Help and Social Closeness*

TWITTER We analyzed average and median response time for questions and the social connection between question asker a and responder r (using $\max(M_r^a, M_a^r)$). For this part of the analysis, we only considered a random sample of 550,680 replies to questions for performance reasons. A linear regression model explains the response time as $-43.55 \cdot \max(M_r^a, M_a^r) + 10,786$ with $p = 0.00$, but does not explain the variance ($R^2 = 0.0007$). We estimated the stability of the result by running the same experiment on a smaller dataset (50,000 replies), where we got comparable results (-47.691, intercept 8,892, $p = 0.00$, $R^2 = 0.0007$). Given the high number of replies and the low p-value, we do not expect the result to change significantly when analyzing a larger subset. User pairs who exchanged a direct message for the first time when answering the question under consideration (i.e., the number of exchanged messages equals to 1) have a high average response time of 3.8 hours (13,791 seconds, SD: 73,795) whereas users who have exchanged at least 1 message before have an average response time of 2.5 hours (9,096 seconds, SD: 54,086). In addition, 90% of the question asker/responder pairs have exchanged $\leqslant 35$ messages. While pairs with no previous interaction have a median response time of 10 minutes (598 seconds), pairs who have directly communicated before have a lower median response time of 7 minutes (420 seconds).

10.7 LIMITATIONS

The interpretation of Facebook's *like* statement as relevance judgment for replies to questions is not optimal, since users do not necessarily associate it with a judgment on content quality. To the same extent, mapping retweets to relevance judgments may e.g. not cover content that is considered as highly interesting by the user but assumedly not relevant for the user's followers and therefore not retweeted. One might also

doubt whether response time is a valid proxy for the willingness to help others – it could as well be the case that people who reply faster to questions received via Twitter do so because they spend a much larger part of their life online and therefore have more and deeper relationships on Twitter. The datasets suffer from high variance, making it difficult to show indications and trends. We only considered direct social relationships, i.e. only a single step within the social graph. In a more sophisticated modeling approach, indirect relationships could also be taken into account.

11

EXPERIMENT 2: INFORMATION SEEKERS' WILLINGNESS TO USE SMQA

11.1 SYNOPSIS

One form of social search is to integrate one's social network in the search process by querying friends, leading to more subjective but also highly individualized answers. Previous studies analyzed users' social search behavior employing (broadcasted) status messages on social networking platforms to communicate information needs (Status Message Question Asking, SMQA, cf. Section 2.5.2) and revealed a limited willingness of information seekers to use SMQA when comparing it to traditional search engines. We describe the results of a survey with 112 participants and show that directly approaching well chosen friends is considered more attractive and is associated with higher expectations in terms of response quality than SMQA. Our findings suggest that users anticipate quality improvements gained from forwarding queries especially for certain content types of information needs and that response time is an important factor for the information seeker.

11.2 MOTIVATION

Previous results (Oeldorf-Hirsch et al., 2014; Morris et al., 2010a) suggest that there is only limited willingness to leverage one's social network to satisfy information needs. In this study, we investigate whether the behavior of information seekers and the expected quality of the response changes when people ask others directly instead of broadcasting the query on a social networking platform. Furthermore, we want to understand which expectations exist in terms of time constraints and whether forwarding the query to others outside the own social network is associated with improved response quality.

A laboratory experiment conducted by Oeldorf-Hirsch et al. (Oeldorf-Hirsch et al., 2014) comparing user's preference to route information needs to a traditional search engine or a social network for SMQA (Facebook or Twitter) revealed that more than a quarter of the participants (27%) did not send any questions to Twitter or Facebook at all. Only $20\% - 24\%$ of the prompted information needs were routed (not exclusively) to SMQA. Recommendations, opinions, and factual knowledge appear to be common candidates for SMQA. Navigational and exploratory information needs were routed almost exclusively to search engines. While the navigational result could have been expected since traditional search engines are known to excel at answering navigational queries, it is surprising that the same seems to apply to exploratory search, a field which has always been considered as a weakness of traditional search engines (White et al., 2006). The main motives for choosing SMQA as a means to answer the information need are in general consistent with the findings of Morris et al. (Morris et al., 2010b) and include higher trust in responses, higher adequateness when asking for subjective information, the assumption of the presence of a specific audience

and better personalization and contextualization. The main reasons why questions did not get posted to a social network are (1) the wrong level of specificity of the information need (either too specific or not specific enough), (2) the perceived lack of available knowledge within one's social network, and (3) the fear of disrupting one's social network (Zhao et al., 2013). Participants of Oeldorf-Hirsch et al. (Oeldorf-Hirsch et al., 2014) indicated that they did not choose to route the information need to their social network because they would like to get a fast reply, whereas Morris et al. (Morris et al., 2010b) reported in contrast that participants intentionally chose SMQA because of the higher answer speed. In Oeldorf-Hirsch et al.'s study (Oeldorf-Hirsch et al., 2014), 39% of the information needs posted to social networks received an answer (ranging from 1 to 10 answers, median of 3). The median response time was 5 hours 55 minutes, with the fastest response 1 hour 34 minutes after the information need was posted on the social network. Having a common context makes asking others more attractive, even for question types which normally fall into the domain of traditional search engines (e.g. factual knowledge in Forte et al.'s study (Forte et al., 2014)). Lampe et al.'s results (Lampe et al., 2014) suggest that the information providers in general are willing to help and therefore treat requests differently than normal status messages.

11.3 RESEARCH QUESTIONS

With the aforementioned vision of a distributed social search system where information seekers query other people's information spaces (cf. Part II), we limit ourselves to the social aspects of manually asking others for information as a first step. We focus on the following questions, which – to the best of our knowledge – are not covered by previous SMQA studies, to continue the efforts to understand the social elements of the human information seeking process:

I. Does the information seeker's preference to ask others increase when the query is not broadcasted among her friends but sent directly to certain designated, potentially knowledgeable contacts within the social network?
II. Does the information seeker expect a higher response quality when the request is not broadcasted but directly sent to potentially knowledgeable contacts?
III. To which degree does forwarding requests to the extended social network (i.e. friends of friends) increase the expected quality of the responses?

RQ I addresses Oeldorf-Hirsch et al.'s finding (Oeldorf-Hirsch et al., 2014) that people are reluctant to post queries to their social network visible to all friends – we would like to understand whether this behavior can get mitigated by framing the audience accordingly. A positive result would suggest that the routing of questions in a social information retrieval scenario not only has a technical component (i.e., who is able to answer a question?) but also a highly social one (i.e., does the information seeker feel comfortable to reveal the information need to a set of potential information providers?). RQ II investigates whether information seekers associate approaching possible experts directly with a higher response quality. Our hypothesis is that by asking potential knowledgeable contacts directly, the fear of disrupting one's social network and the limited specificity as mentioned by Oeldorf-Hirsch et al. (Oeldorf-Hirsch et al., 2014) can get mitigated. RQ III investigates whether the

assumed limitation of available knowledge within the social network mentioned in (Oeldorf-Hirsch et al., 2014) can get resolved by including the extended social network in the search process.

11.4 STUDY SETUP

For better comparability with existing literature we used the information needs assembled by Oeldorf-Hirsch et al. (Oeldorf-Hirsch et al., 2014) as a basic component of our study (Table 6). The exhaustive set consists of 30 information needs clustered in 10 topics (3 information needs per topic cluster) with the objective to cover a broad set of different areas. We translated the information needs to German to avoid any confusion evoked by the fact that the participants are not native English speakers and removed any references to a laboratory study (e.g. "Find a good place where you can get some food *after this study*" has been changed to "Find a good place (...) some food *today*").

The survey had two parts. In the first part, participants decided how they would satisfy 10 information needs taken from Oeldorf-Hirsch et al.'s list (Oeldorf-Hirsch et al., 2014) (Table 6). The information needs covered all 10 topic clusters, so for each of the 10 topic clusters one information need was selected randomly. For each information need the possible answer options were

- Search using a search engine (e.g. Google, Bing),
- Post a question on a social networking platform (e.g. Facebook, Google+, Twitter) visible to all your friends, and
- Ask individually selected friends (e.g. via email, phone, face-to-face, messaging services).

As in Oeldorf-Hirsch et al.'s study (Oeldorf-Hirsch et al., 2014), selecting multiple options was possible. If the participant chose at least one of the social options (option 2 and 3) she was asked to specify the maximum acceptable response time from the social network or the selected friends (possible values: $<$ 1 hour, 1-5 hours, 5-24 hours, $>$ 24 hours). In the second part of the survey, participants estimated the quality of the answers they would expect to receive when sending the question to (1) their social network for SMQA or (2) their well chosen friends. Participants were asked to estimate whether a reply would answer the information need (a) not at all; (b) not really, but increases the understanding; (c) partially; (d) mostly; (e) completely for 3 randomly selected information needs. In addition, we asked to which degree forwarding the question to others would improve the quality of the answer (possible options: not at all; low; medium; high).

Recommendation
Find a good place (restaurant, diner/takeaway etc.) where you can get some food today
Find a good birthday present for a specific relative
Imagine a trip you would like to take in the future and find out what others recommend as the best sights to see
Opinion

Think of a certain place you are interested in seeing and find out whether it's worth traveling there
Think of the next tech product you'd like to buy and find out what people think of it
Think of a TV show that you plan to watch during your next free hour and find out what others think of the show
Factual knowledge
Find out what might be causing symptoms you have been having recently
Find out what traffic will be like for your commute to your next destination
Find out what the weather is like outside right now
Rhetorical
Contemplate something that's always confused you - see what others think
Think of something that's frustrating you right now - see what others think
Think of a strong opinion you have about a current issue - see what others think
Invitation
Plan an activity you would like to do this weekend and find out who is interested in joining you
Find out if someone in the area is interested in meeting up for your next meal
Think of something you would like to do tomorrow and find out if anyone else would be interested too
Favor
Think of a project you'd like to do or a task you need to finish for which you do not have the right tool or gadget. Find someone local who has this particular item you can borrow
Think of a task at home you could use help with today, and find someone who would be willing and available to help
Think of an errand that needs to get done today. Find someone else who can take care of it right now
Social Connection
Find someone who can help you learn more about a new hobby you'd like to take up
Find someone who would be a good person to know for finding a job in your local city for you or someone else
Find someone to teach you a new skill while you are online right now
Offer
Think of a skill or particular area of knowledge you have. Find someone who could benefit from this skill/knowledge
Think of an item you have at home that you no longer use. Find someone else who could use it
Think of something you can offer to do in your next free hour that would be useful within your group of friends or local community
Navigational
Find the website for the main gym in your neighborhood
Find Nike's website
Find website of local library
Undirected / Exploratory
Find a current event you are interested in keeping up with (one you are not already keeping up with)
Find an idea for a new hobby (a hobby you haven't considered before)
Find a new activity to do this week

Table 6: Information needs assembled by Oeldorf-Hirsch et al. (Oeldorf-Hirsch et al., 2014) (slightly adjusted to remove references to a laboratory study; participants received questions translated to German)

11.5 PARTICIPANTS

The survey was conducted online with a group of 112 participants (36% female). The participants have been recruited online and received a small compensation for their time (< 10 EUR). Out of the 112 participants, 57 (51%) are students, 50 (45%) are employed in a variety of jobs and 5 (5%) are either trainees, pensioners or homemakers. All of them live in Germany. The average age is 25.5 years (SD: 9.1). On average, each participant has 220.3 friends on social networking platforms (SD: 233.1). While the participant group is not necessarily representative for the population of an average country one could argue that the group can be regarded as target group for modern information seeking approaches and it therefore is suitable to infer conclusions from the preferences of the group members.

Following the method explained above each participant routed 10 information needs to either a search engine, a status message in a social network platform (SMQA) or a friend (multiple selections were possible), depending on the individual preference to solve the respective information need (summing up to 1,120 routing decisions in total). The participants did only indicate which path they would choose, they were not asked to conduct a web search, post something on a social network, or send a message to a friend. Afterwards, each participant rated the expected quality of the reply and the assumed quality gain when forwarding the message to the extended social network for 3 randomly selected information needs (leading to 336 assessments in total).

11.6 RESULTS

11.6.1 *RQ I: Preference to Involve Others to Satisfy Information Needs*

Out of 112 participants, 30 (27%) would not post anything to their social network platform while only 6 (5%) chose not to ask a single question to one of their friends.

Out of 1,120 routing decisions, 763 (68%) have been routed to a search engine, 536 (48%) to a friend and 256 (23%) to a status message on a social networking platform. Figure 6 shows the split by type of information need. Not surprisingly, search engines have mostly been selected for information needs with navigational, factual knowledge, recommendation, opinion, and undirected/exploratory types. Asking friends directly was primarily chosen for information needs within the categories favor, invitation, offer, and social connection. While 472 (42%) of all routing decisions were solely routed to search engines, 357 (32%) were addressed to social targets exclusively (directly to friends only: 208/19%; to SMQA only: 68/6%; to both: 81/7%) and 291 (26%) were sent to search engines and social targets (search engines and directly to friends: 184/16%; search engines and SMQA: 44/4%; search engines, friends directly and SMQA: 63/6%).

11.6.2 *RQ II: Expected Response Quality*

We asked each participant to estimate the expected response quality for 3 information needs on a 5-point scale (5=best) when (a) asking well chosen friends individually or

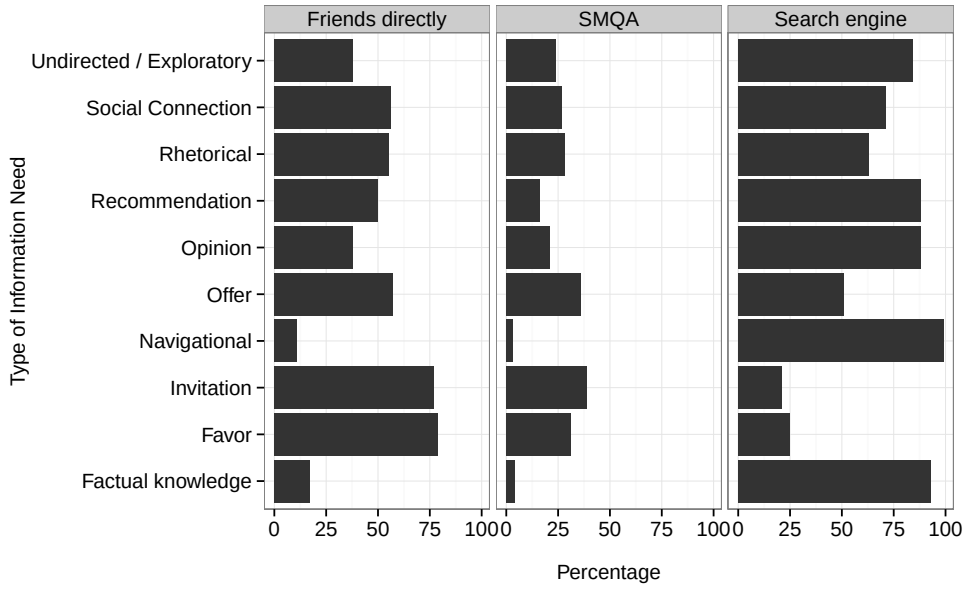


Figure 6: Routing decisions of participants by type of information need (Experiment 2)

(b) using SMQA on a social network platform. On average, participants rated the expected quality of a potential response received from a directly asked friend with 3.86 (SD: 1.04), whereas responses from SMQA received an average rating of 3.10 (SD: 1.18). Conducting a 2-sample *paired t-test* (with "quality" as dependent variable) confirmed that the results are statistically significant ($p = 0.00$, $df = 335$, $t = 11.17$).

11.6.3 *RQ III: Forwarding of Requests*

We also asked the participants to assume that a friend already forwarded their query to her friends and to rate on a 4-point scale to which degree this would improve the quality of the received answers (4=best). The overall expectation was that it would have a low to medium effect, with the biggest benefit for information needs in the categories exploratory (∅2.9, SD: 0.7), offer (∅2.9, SD: 0.9), opinion (∅2.8, SD: 0.9), and recommendation (∅2.8, SD: 0.8). The lowest average quality increase was anticipated for information needs belonging to the categories factual knowledge (∅2.3, SD: 1.0), favor (∅2.6, SD: 1.0), and navigational (∅2.6, SD: 0.9).

11.7 DISCUSSION AND LIMITATIONS

11.7.1 *RQ I: Preference to Involve Others to Satisfy Information Needs*

Figure 7 illustrates the distribution of routing decisions and compares Oeldorf-Hirsch et al.'s findings (Oeldorf-Hirsch et al., 2014) with our results. When interpreting the data, one needs to bear in mind that in contrast to Oeldorf-Hirsch et al. (Oeldorf-Hirsch et al., 2014), we did not conduct a laboratory experiment but gained the data from an online survey. Oeldorf-Hirsch et al. posted a subset of the information needs to social networking platforms, while we only asked the participants for their routing

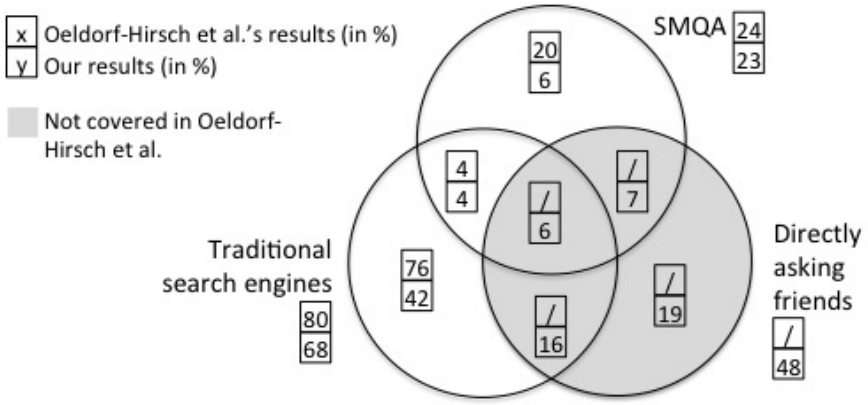


Figure 7: Participants' preferences to satisfy information needs in Experiment 2 (comparison with results from Oeldorf-Hirsch et al. (Oeldorf-Hirsch et al., 2014))

decisions. Since some key figures are of similar size (e.g. SMQA ratio of information needs, ratio of people who do not want to post anything on a social networking platform) we are quite confident that the survey data reflects the participants' preferences in a meaningful way.

Looking at SMQA alone, our results confirm the outcome of Oeldorf-Hirsch et al. (Oeldorf-Hirsch et al., 2014) (24% in Oeldorf-Hirsch et al. vs. 23% in our study). The results also show that people are willing to ask others for help when this can be done directly (76% of the information needs in Oeldorf-Hirsch et al.'s study (Oeldorf-Hirsch et al., 2014) were routed to traditional search engines exclusively – when offering the option to ask others directly, this percentage drops to 42%). The same chain of reasoning applies when looking at the participants: While 27% of the participants did not route any information need to social networks in both studies, only 5% of our participants chose not to ask a single question to a friend. To leverage the potential of a social network for search, it is important to allow asking users directly (and not only sending broadcasts).

11.7.2 *RQ II: Expected Response Quality*

Since we did not post anything to a social network or send requests to contacts within the information seeker's social network, it is not possible to compare these results with previous studies. As already mentioned above, asking others directly is associated with a higher response quality than SMQA. This is in line with previous findings of Teevan et al. (Teevan et al., 2011).

11.7.3 *RQ III: Forwarding of Requests*

Forwarding of requests was assumed to be beneficial for certain types of information needs (exploratory, offer, opinion, and recommendation) which could possibly be provided by people with a higher social distance – therefore, it might help to include them. Information needs where information seekers offer something to other people might not cause social discomfort when they are asked to someone not known in

person. On the contrary, forwarding was not considered useful for information needs of the categories factual knowledge, favor and navigational – the participants might have been hesitant to ask strangers for things they could easily look up themselves (factual knowledge, navigational) and might not have assumed that strangers would do them a favor. The outcome suggests that including a forwarding mechanism in a social search engine would be useful for certain types of information needs and makes also sense from a network theory perspective ((Milgram, 1967)).

12

EXPERIMENT 3: CLASSIFICATION OF INFORMATION NEEDS FOR SOCIAL INFORMATION RETRIEVAL

The following experiment has been conducted in the context of Ruth Nussbaumer's Bachelor's Thesis (Nussbaumer, 2014) (April to July 2014) and Akash Nayyar's Master's Thesis (Nayyar, 2015) (May to November 2015), both of which were supervised by Christoph Fuchs and Georg Groh at the Chair for Applied Informatics – Cooperative Systems at Technische Universität München.

12.1 SYNOPSIS

Some information needs might be better suited to get satisfied using social information retrieval techniques than others because of inherent features of the information need. In the following section, we describe an experiment where prominent websites from various content categories are used to represent their respective content area and allow to correlate attributes of the content areas. The underlying assumption is that successful websites for focused content areas perfectly align with the information seekers' requirements when satisfying information needs in the respective content areas. Based on a manually collected dataset of URLs from websites covering a broad range of topics taken from Alexa[1] (a company that publishes statistics about web traffic), a crowdsourcing approach is employed to rate the information needs that could get solved by the respective URLs according to several dimensions (incl. sociality and mobility) to investigate possible correlations with other attributes. Our results suggest that information needs which do not require a certain formal expertise play an important role in social information retrieval and that some content areas are better suited for social information retrieval (e.g., Factual Knowledge & News, Games, Lifestyle) than others (e.g., Health & Lifestyle).

12.2 MOTIVATION

With an increasing number of available low-cost capabilities to sense the user's individual environment it becomes possible to grasp and consider the user's context (cf. Section 2.4) in search situations. In the following experiment, we investigate whether information needs covering content areas with different characteristics benefit from context-awareness and/or social means to satisfy information needs.

12.3 RESEARCH QUESTIONS

The objective of this experiment is to investigate whether there are types of information needs (either on a meta-level or content-wise) which are more "mobile" or

1 http://www.alexa.com (retrieved 2015-11-04)

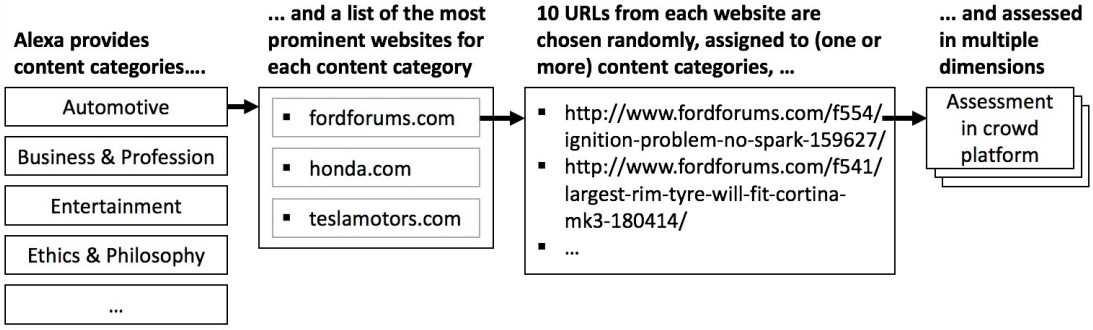


Figure 8: Study approach (Experiment 3)

"social" than others and whether specific attributes can predict or explain the "sociality" of an information need.

Mobile information needs are included in the analysis to evaluate the hypothesis that mobile information needs show similar characteristics as information needs profiting from sociality (as already suspected by (Church and Oliver, 2011)).

12.4 APPROACH

To structure information needs by content type, a content taxonomy provided by Alexa, a company that publishes information about web traffic (and is owned by Amazon), is used. Alexa also issues a list of the most popular websites for each category. Alexa estimates the popularity of a URL by tracking a subset of the web users with the Alexa toolbar, a plugin for web browsers. The approach to answer the research questions of this experiment is the following: the basic assumption is that the most prominent websites for each content category are successful because they satisfy the users' information needs (which brought the users to the website in the first place) in that specific content area in an adequate way. The successful websites offer functionalities and features that fit to the content area, so that people who engage with the content feel comfortable with it. Instead of analyzing the information needs that might occur in the different categories, we investigate the websites which satisfy the information needs in each content category (see next section), assuming that the websites "respond" to the queries in the most appropriate way.

For each Alexa content category, the three most prominent websites have been selected (only exceptions: Travel and Ethics & Philosophy, which only consist of two websites). Since some websites cover topics which can not be linked to a single content dimension, the topic of each website is expressed as a vector in Alexa's content category vector space holding percentage values for each content dimension. For each website, ten randomly chosen URLs have been selected and rated by participants of a web survey in different dimensions explained in the next section (Section 12.5). Figure 8 gives an overview of the approach. A full list of categories and websites is listed in the appendix (Table 36). The participants of the survey received a small compensation for the task. We only accepted those judgments where the elapsed time between showing and submitting the survey form suggests that the user read and understood the questions (Section 12.5.3).

12.5 DATASET

Each URL was rated on several dimensions using a web survey. The dimensions can be divided in two groups: the first group (Section 12.5.1) is related to the hypothetical information need which would cause someone to visit the URL, the second group (Section 12.5.2) discusses the website's direct properties (business model, types of fostered social interaction, etc.). We do not claim that the dimensions are collectively exhaustive – others do exist (cf. Section 12.7 for examples), but we focused on the ones listed in the next section because we assumed that those might have a high likelihood of showing differences between information needs with high and low levels of sociality.

12.5.1 *Dimensions to Classify Information Needs*

In the following, the dimensions to classify the information need that is satisfied on the respective URL are explained in detail.

DEPENDENCE ON TIME This dimension reflects whether the information need or the information have a certain expiration date or not. Possible values are

- HARD CONSTRAINT (2): The information need needs to be addressed at a specific point in time, e.g., if a user needs to get an idea for a good Christmas present, then the information need is clearly addressed at a specific time with the urgency depending on the current date.
- SOFT CONSTRAINT (1): The information need is addressed at a vague time, e.g., the information need is of the kind: *"Any ideas for summer holidays?"*.
- INDEPENDENT (0): Not dependent on the time, e.g., the information need is of the kind: *"What is your favorite football team?"*.

TEMPORAL VALIDITY This dimension describes how long the information presented is valid. Possible values are

- LONG (2): Reply to information need is valid for a very long time (e.g., decades, centuries, forever) , e.g., *"When was Mozart born?"* – the answer is valid forever.
- MEDIUM (1): Reply to information need is valid for a long time (e.g., months, maybe years), e.g., *"Who is the current football player of the year?"*, or *"How much does the new MacBook Pro cost?"*
- SHORT (0): Reply to information need is valid for a short time (e.g. hours, days, maybe weeks). For instance, *"Will it be sunny tomorrow?"*

GENERAL APPLICABILITY AMONG USERS This dimension describes to which degree the knowledge that is presented (and searched for in the hypothetical information need) is applicable for multiple users. Possible values are

- HIGH (2): Information needs that are not tailored to one particular user; information need is shared by a lot of people, for instance: *"What is the average cost of living in Munich?"*
- MEDIUM (1): Information needs that are important for a specific subset of people, for instance: *"How to compute running time in Java?"*. In this case the given information need targets programmers.
- LOW (0): Information needs that are only important for a particular user, for instance: *"My GRE score is 310. Is this sufficient to get admission in a good university in the United States?"*. In this case the information need is related only to the specific user who has a GRE score of 310.

KNOWLEDGE CODIFICATION This dimension expresses to which degree the knowledge that is required to satisfy the information need is codified. In mature fields with commonly accepted explicit forms of knowledge representation (e.g., books, websites, etc.), knowledge is codified to a higher degree (e.g., medicine) than in areas where knowledge is widely discussed and controversial (e.g., user experience with the new BMW i3). Possible values are

- HIGH (1): Knowledge to satisfy the information need is codified, i.e., it is defined in form of facts (books or articles) and there is a common agreement, e.g.: *"What are the symptoms of Parkinson's disease?"*
- LOW (0): Knowledge to satisfy the information need is not codified, i.e., it is widely discussed and controversial; examples include questions asking for recommendation (e.g., *"What is the best restaurant in Munich?"*) or opinions (e.g., *"Do you like the new BMW i3?"*)

12.5.2 *Dimensions to Classify Specialized Websites*

COSTS The cost dimension describes whether the access to the information on the website is free or requires payment. Possible values are FEE (2), PARTIALLY FREE (1) and FREE (0).

INFORMATION PROVIDER This attribute describes which profile fits best to the person who provides the information on the respective URL. Possible values are EXPERT (e.g., doctor, lawyer, editor), OPERATOR (e.g., someone informing about own services or products), and LAYMAN (i.e., someone who not necessarily has any formal expertise on the subject).

SOCIALITY For each website, the existence of the following social features is evaluated and forms the degree of sociality (all features are weighted equally):

- Can an ordinary user ask questions to satisfy an information need?
- Does the website recommend other content that was liked, commented on, viewed, or posted by others?

- Is there a possibility to rate or comment on the information need?
- Is there a possibility to create a personal profile?
- Is it possible to see what kind of information needs other people have or what kind of information needs they have satisfied before?
- Is it possible to contact the user who had the information need?

MOBILITY This dimension consists of three equally weighted sub-dimensions (aggregated to a single mobility value) and describes to which degree the underlying information need represents a "typical" mobile scenario. The participant is asked whether the information need which is satisfied by the given URL depends on a specific location. Valid answers include

- HIGH: The user's physical location has a definite impact on the information need and the type of answer expected, e.g.: *"Where is the ALDI supermarket closest to Klinikum Grosshadern metro station in Munich, Germany?"*
- LOW: The user's location does not impact the information need, for instance: *"Which is the best Android phone in the market at present?"*

In addition, the participant is asked whether it is likely that the information need occurred in a mobile context and whether the information contains any specific spatial location information.

12.5.3 *Data Collection*

Data collection was conducted using an online survey on a crowdsourcing platform[2] with Indian participants. Each participant assessed ten randomly chosen URLs, using the dimensions outlined above. For each website, the data of the related URLs was aggregated using the average of the respective URL ratings and normalized on the interval $[0, 1]$. To ensure data quality, all submissions that took less than 60 seconds were excluded from the evaluation and were added to the pool of untreated URLs again. The threshold of 60 seconds has been identified using test runs with skilled English speakers. In total, the dataset used for the analysis consists of 532 evaluated URLs taken from 52 websites.

12.6 RESULTS

12.6.1 *Correlations*

The correlation of the different content categories and dimensions is shown in Table 7 (Spearman's rho) and Table 8 (Pearson's r). In the following, the findings will be briefly discussed for each dimension.

2 `https://microworkers.com/` (retrieved 2016-01-12)

DEPENDENCE ON TIME Dependence on Time is positively correlated with Business Profession, Automobiles, Health & Lifestyle, and Entertainment. While the first two content categories could possibly be explained by pressing information needs before (purchasing) decisions, the relation for the last two is less obvious. The dimension Layman is negatively correlated, which intuitively makes sense when considering that normal users will not be the best information providers when time critical information is requested. Content categories with a high negative correlation with dependence on time are Sports, Games, Real Estate, and Society. Especially for the categories Sports and Real Estate, this result is surprising. Real Estate refers to renting or buying a property and the result might indicate that these decisions are rather short-dated than initially assumed.

TEMPORAL VALIDITY Content in the category Ethics & Philosophy positively correlates with Temporal Validity. This is not surprising, since the content is not expected to change fast. In contrast, content in the areas Entertainment, Sports, and Technology varies at a much higher pace and therefore is negatively correlated.

GENERAL APPLICABILITY On average, people's information needs regarding Travel do not seem to differ much, since information in the Travel category is positively correlated with General Applicability. The findings suggest that the same applies to Business Profession and Health & Lifestyle. In contrast, topics like Society, Ethics & Philosophy, and Lifestyle seem to be discussed quite individually – for the Ethics & Philosophy category this comes a bit unforeseen, however, when taking the discussion and interpretation into account, the result may become more understandable.

KNOWLEDGE CODIFICATION The findings suggest that Business Profession and Recreation tend to have a high degree of knowledge codification; in addition, it is also positively correlated with General Applicability (Spearman's rho only) and Temporal Validity (information that is valid for a long time or that is valid for a large group tend to be codified to a higher degree than other information). On the other end of the spectrum, information in the categories Society, Games, and Sports are negatively correlated with Knowledge Codification. The negative correlation with Factual Knowledge & News is surprising.

COSTS Websites in the content areas Lifestyles, Technology and Automobiles tend to have higher results in the Costs dimension. Also the categories Sociality and Layman are positively correlated with Costs. The Costs dimension is negatively correlated with Ethics & Philosophy, Health & Lifestyles, Trivia, and Finance & Insurance. Especially the last category is unexpected because intuitively people would be willing to invest in serious topics like finance when stakes are high.

LAYMAN Layman has a high positive correlation with Sociality. In addition, it is positively correlated with topics in the categories Society and Automobiles. In contrast, Layman is negatively correlated with Dependence on Time and Operator, which could be caused by the fact that laymen typically need some time to reply to

information needs and the type of interaction. In addition, LAYMAN is negatively correlated with ENTERTAINMENT, which is surprising since one could intuitively assume that "informal" topics are related to less professional interaction modes.

OPERATOR Operator is positively correlated with the content categories HOMES & GARDEN, and SPORTS. A negative correlation exists for the dimensions SOCIALITY, FINANCE & INSURANCE, and SOCIETY.

EXPERT The EXPERT dimension is positively correlated with AUTOMOBILES, GENERAL APPLICABILITY, and REAL ESTATE. It seems valid to assume that the knowledge of experts is applicable to a larger audience and that expensive purchasing decisions might be backed up by acknowledged expertise. The positive correlation with KNOWLEDGE CODIFICATION suggests that experts work in mature, clearly distinguished fields with commonly accepted methods and a documented state of the art. Negatively correlated are SPORTS, TECHNOLOGY, and MOBILITY. The first two categories could be explained by the fact that both are content-driven and expertise might be easier to gain (or maybe difficult to get, because of low degree of knowledge codification as in SPORTS). The negative correlation between EXPERT and MOBILITY could possibly be explained because experts for a certain spatial area are often not considered as "professional" experts and therefore correspond more with the LAYMAN category.

		Dimension										Content category																			
		Dependence on Time	Temporal Validity	General Applicability	Knowledge Codification	Costs	Layman	Operator	Expert	Sociality	Mobility	Entertainment	Automobiles	Finance & Insurance	Food & Drink	Health & Lifestyle	Factual Knowledge & News	Business Profession	Real Estate	Sports	Games	Technology	Travel	Society	Ethics & Philosophy	Recreation	Homes & Garden	Lifestyle	Trivia
Dimension	Dependence on Time	1.00	0.02	0.11	0.13	0.08	-0.22	-0.01	0.04	0.10	0.03	0.10	0.22	0.00	-0.05	0.14	0.11	0.23	-0.17	-0.27	-0.18	0.07	-0.06	-0.11	-0.03	-0.02	0.07	0.03	-0.03
	Temporal Validity	0.02	1.00	0.03	0.18	-0.05	0.22	0.04	-0.01	0.07	0.03	-0.23	0.13	-0.03	-0.02	0.04	0.14	0.19	0.06	-0.15	-0.07	-0.15	0.05	-0.01	0.28	0.11	0.17	-0.13	-0.11
	General Applicability	0.11	0.03	1.00	0.18	0.08	0.02	-0.06	0.18	0.14	-0.05	-0.08	-0.04	-0.07	-0.03	0.16	0.01	0.20	0.01	0.09	0.01	-0.04	0.42	-0.24	-0.12	0.08	-0.03	-0.11	0.11
	Knowledge Codification	0.13	0.18	0.18	1.00	-0.07	-0.04	0.07	0.13	0.04	-0.03	-0.11	0.10	-0.05	-0.10	0.17	-0.13	0.29	-0.06	-0.14	-0.14	0.11	0.14	-0.16	0.07	0.23	0.06	0.06	0.00
	Costs	0.08	-0.05	0.08	-0.07	1.00	0.20	0.04	-0.08	0.21	0.11	0.11	0.17	-0.19	0.10	-0.11	-0.03	0.01	0.02	-0.07	0.06	0.32	-0.03	-0.06	-0.18	0.05	0.14	0.38	-0.17
	Layman	-0.22	0.22	0.02	-0.04	0.20	1.00	-0.28	0.07	0.46	0.25	-0.21	0.27	-0.06	-0.14	-0.22	0.05	-0.13	0.06	0.13	-0.04	0.01	0.09	0.46	0.14	0.10	-0.01	-0.11	0.07
	Operator	-0.01	0.04	-0.06	0.07	0.04	-0.28	1.00	-0.06	-0.35	0.05	-0.04	0.09	-0.21	0.04	-0.12	-0.13	0.02	0.06	0.10	0.06	0.03	0.03	-0.23	0.02	-0.11	0.13	-0.11	-0.05
	Expert	0.04	-0.01	0.18	0.13	-0.08	0.07	-0.06	1.00	0.10	-0.10	0.01	0.25	-0.05	0.03	0.07	-0.05	0.00	0.16	-0.26	0.10	-0.09	-0.01	-0.07	0.11	-0.04	-0.04	-0.05	0.03
	Sociality	0.10	0.07	0.14	0.04	0.21	0.46	-0.35	0.10	1.00	0.16	-0.07	0.09	-0.08	-0.01	-0.18	0.23	0.22	-0.04	-0.08	0.25	-0.03	0.07	0.18	-0.17	0.07	0.14	0.17	-0.06
	Mobility	0.03	0.03	-0.05	-0.03	0.11	0.25	0.05	-0.10	0.16	1.00	-0.07	-0.05	0.09	0.01	-0.16	0.16	0.12	0.08	0.03	0.04	0.25	-0.13	-0.03	0.23	-0.32	-0.05	0.05	0.04
Content category	Entertainment	0.10	-0.23	-0.08	-0.11	0.11	-0.21	-0.04	0.01	-0.07	-0.07	1.00	-0.11	-0.16	-0.13	-0.11	-0.13	-0.16	-0.15	-0.11	0.32	0.11	-0.15	-0.02	-0.09	-0.09	-0.13	0.09	-0.11
	Automobiles	0.22	0.13	-0.04	0.10	0.17	0.27	0.09	0.25	0.09	-0.05	-0.11	1.00	-0.09	-0.07	-0.06	-0.07	-0.09	-0.08	-0.06	-0.06	0.16	-0.08	0.21	-0.05	-0.11	-0.07	-0.09	-0.06
	Finance & Insurance	0.00	-0.03	-0.07	-0.05	-0.19	-0.06	-0.21	-0.05	-0.08	0.09	-0.16	-0.09	1.00	-0.10	-0.09	0.52	0.04	0.30	-0.09	-0.09	0.16	-0.12	-0.14	-0.07	-0.15	-0.10	-0.13	-0.09
	Food & Drink	-0.05	-0.02	-0.03	-0.10	0.10	-0.14	0.04	0.03	-0.01	0.01	-0.13	-0.07	-0.10	1.00	0.23	-0.08	-0.10	-0.09	-0.07	-0.07	-0.11	-0.09	-0.02	-0.06	-0.12	-0.08	0.13	-0.07
	Health & Lifestyle	0.14	0.04	0.16	0.17	-0.11	-0.22	-0.12	0.07	-0.18	-0.16	-0.11	-0.06	-0.09	0.23	1.00	-0.07	-0.09	-0.08	-0.06	-0.06	-0.10	-0.08	0.02	-0.05	-0.11	-0.07	-0.09	-0.06
	Factual Knowledge & News	0.11	0.14	0.01	-0.13	-0.03	0.05	-0.13	-0.05	0.23	0.16	-0.13	-0.07	0.52	-0.08	-0.07	1.00	0.10	-0.09	-0.07	-0.07	0.26	-0.09	0.10	-0.06	-0.12	-0.08	-0.10	-0.07
	Business Profession	0.23	0.19	0.20	0.29	0.01	-0.13	0.02	0.00	0.22	0.12	-0.16	-0.09	0.04	-0.10	-0.09	0.10	1.00	0.29	-0.09	-0.09	0.01	-0.12	-0.23	-0.07	-0.15	0.09	0.07	-0.09
	Real Estate	-0.17	0.06	0.01	-0.06	0.02	0.06	0.06	0.16	-0.04	0.08	-0.15	-0.08	0.30	-0.09	-0.08	-0.09	0.29	1.00	-0.08	-0.08	-0.13	-0.11	-0.20	-0.07	-0.14	-0.09	-0.12	-0.08
	Sports	-0.27	-0.15	0.09	-0.14	-0.07	0.13	0.10	-0.26	-0.08	0.03	-0.11	-0.06	-0.09	-0.07	-0.06	-0.07	-0.09	-0.08	1.00	-0.06	-0.10	-0.08	0.20	-0.05	-0.11	-0.07	-0.09	-0.06
	Games	-0.18	-0.07	0.01	-0.14	0.06	-0.04	0.06	0.10	0.25	0.04	0.32	-0.06	-0.09	-0.07	-0.06	-0.07	-0.09	-0.08	-0.06	1.00	-0.10	-0.08	-0.16	-0.05	-0.11	-0.07	-0.09	-0.06
	Technology	0.07	-0.15	-0.04	0.11	0.32	0.01	0.03	-0.09	-0.03	0.25	0.11	0.16	0.16	-0.11	-0.10	0.26	0.01	-0.13	-0.10	-0.10	1.00	-0.13	-0.08	-0.08	-0.17	0.08	-0.14	-0.10
	Travel	-0.06	0.05	0.42	0.14	-0.03	0.09	0.03	-0.01	0.07	-0.13	-0.15	-0.08	-0.12	-0.09	-0.08	-0.09	-0.12	-0.11	-0.08	-0.08	-0.13	1.00	-0.20	-0.07	0.60	-0.09	-0.12	0.19
	Society	-0.11	-0.01	-0.24	-0.16	-0.06	0.46	-0.23	-0.07	0.18	-0.03	-0.02	0.21	-0.14	-0.02	0.02	0.10	-0.23	-0.20	0.20	-0.16	-0.08	-0.20	1.00	0.09	-0.15	-0.18	-0.13	-0.16
	Ethics & Philosophy	-0.03	0.28	-0.12	0.07	-0.18	0.14	0.02	0.11	-0.17	0.23	-0.09	-0.05	-0.07	-0.06	-0.05	-0.06	-0.07	-0.07	-0.05	-0.05	-0.08	-0.07	0.09	1.00	-0.09	-0.06	-0.07	-0.05
	Recreation	-0.02	0.11	0.08	0.23	0.05	0.10	-0.11	-0.04	0.07	-0.32	-0.09	-0.11	-0.15	-0.12	-0.11	-0.12	-0.15	-0.14	-0.11	-0.11	-0.17	0.60	-0.15	-0.09	1.00	0.08	0.16	-0.11
	Homes & Garden	0.07	0.17	-0.03	0.06	0.14	-0.01	0.13	-0.04	0.14	-0.05	-0.13	-0.07	-0.10	-0.08	-0.07	-0.08	0.09	-0.09	-0.07	-0.07	0.08	-0.09	-0.18	-0.06	0.08	1.00	0.13	-0.07
	Lifestyle	0.03	-0.13	-0.11	0.06	0.38	-0.11	-0.11	-0.05	0.17	0.05	0.09	-0.09	-0.13	0.13	-0.09	-0.10	0.07	-0.12	-0.09	-0.09	-0.14	-0.12	-0.13	-0.07	0.16	0.13	1.00	-0.09
	Trivia	-0.03	-0.11	0.11	0.00	-0.17	0.07	-0.05	0.03	-0.06	0.04	-0.11	-0.06	-0.09	-0.07	-0.06	-0.07	-0.09	-0.08	-0.06	-0.06	-0.10	0.19	-0.16	-0.05	-0.11	-0.07	-0.09	1.00

Table 7: Correlation between dimensions and content categories of information needs (Experiment 3, Spearman's rho)

		Dimension										Content category																			
		Dependence on Time	Temporal Validity	General Applicability	Knowledge Codification	Costs	Layman	Operator	Expert	Sociality	Mobility	Entertainment	Automobiles	Finance & Insurance	Food & Drink	Health & Lifestyle	Factual Knowledge & News	Business Profession	Real Estate	Sports	Games	Technology	Travel	Society	Ethics & Philosophy	Recreation	Homes & Garden	Lifestyle	Trivia
Dimension	Dependence on Time	1.00	0.00	0.08	0.11	0.01	-0.31	0.03	0.07	0.04	-0.05	0.18	0.17	0.00	0.03	0.16	0.09	0.26	-0.12	-0.19	-0.25	-0.03	-0.03	-0.13	-0.08	-0.01	-0.03	0.03	0.01
	Temporal Validity	0.00	1.00	0.10	0.22	-0.13	0.25	0.04	-0.01	0.08	0.06	-0.26	0.15	-0.08	0.00	0.11	0.12	0.17	0.09	-0.10	-0.10	-0.20	0.09	-0.05	0.27	0.14	0.16	-0.11	-0.14
	General Applicability	0.08	0.10	1.00	0.07	0.07	0.08	-0.04	0.17	0.06	-0.08	-0.09	0.00	-0.13	-0.01	0.18	0.03	0.22	-0.01	0.07	0.01	-0.02	0.41	-0.29	-0.15	0.08	0.01	-0.18	0.07
	Knowledge Codification	0.11	0.22	0.07	1.00	-0.06	-0.05	0.06	0.15	0.00	-0.05	-0.29	0.13	-0.01	-0.06	0.16	-0.09	0.24	-0.09	-0.09	-0.29	0.12	0.15	-0.16	0.18	0.22	0.01	0.03	0.04
	Costs	0.01	-0.13	0.07	-0.06	1.00	0.14	0.11	-0.13	0.10	0.14	0.06	0.09	-0.21	0.02	-0.13	-0.08	-0.05	0.11	-0.09	-0.02	0.44	-0.09	-0.15	-0.13	0.00	0.08	0.43	-0.18
	Layman	-0.31	0.25	0.08	-0.05	0.14	1.00	-0.26	0.00	0.44	0.29	-0.21	0.29	-0.16	-0.14	-0.28	0.07	-0.15	0.08	0.05	-0.03	-0.07	0.08	0.38	0.08	0.07	-0.05	-0.08	0.08
	Operator	0.03	0.04	-0.04	0.06	0.11	-0.26	1.00	-0.04	-0.31	0.11	-0.05	0.09	-0.18	0.17	-0.09	-0.10	0.00	0.11	0.15	0.03	0.11	0.00	-0.21	0.11	-0.13	0.27	-0.08	-0.09
	Expert	0.07	-0.01	0.17	0.15	-0.13	0.00	-0.04	1.00	0.09	-0.11	0.00	0.22	-0.02	-0.01	0.13	-0.02	-0.04	0.12	-0.20	0.10	-0.14	-0.01	-0.08	0.06	-0.03	-0.03	-0.02	0.07
	Sociality	0.04	0.08	0.06	0.00	0.10	0.44	-0.31	0.09	1.00	0.18	-0.13	0.05	-0.16	-0.04	-0.31	0.30	0.16	-0.04	-0.11	0.20	-0.07	0.08	0.14	-0.18	0.07	0.07	0.13	-0.02
	Mobility	-0.05	0.06	-0.08	-0.05	0.14	0.29	0.11	-0.11	0.18	1.00	-0.09	-0.06	0.07	0.06	-0.15	0.13	-0.05	0.05	0.08	0.12	0.17	-0.12	-0.05	0.27	-0.30	-0.12	0.03	0.06
Content category	Entertainment	0.18	-0.26	-0.09	-0.29	0.06	-0.21	-0.05	0.00	-0.13	-0.09	1.00	-0.10	-0.13	-0.11	-0.10	-0.11	-0.14	-0.13	-0.10	0.15	0.02	-0.13	-0.09	-0.08	-0.12	-0.11	-0.02	-0.10
	Automobiles	0.17	0.15	0.00	0.13	0.09	0.29	0.09	0.22	0.05	-0.06	-0.10	1.00	-0.08	-0.07	-0.06	-0.07	-0.08	-0.08	-0.06	-0.06	0.14	-0.08	0.17	-0.05	-0.10	-0.07	-0.09	-0.06
	Finance & Insurance	0.00	-0.08	-0.13	-0.01	-0.21	-0.16	-0.18	-0.02	-0.16	0.07	-0.13	-0.08	1.00	-0.09	-0.07	0.23	-0.04	0.19	-0.07	-0.07	-0.01	-0.10	-0.16	-0.06	-0.13	-0.08	-0.11	-0.08
	Food & Drink	0.03	0.00	-0.01	-0.06	0.02	-0.14	0.17	-0.01	-0.04	0.06	-0.11	-0.07	-0.09	1.00	0.10	-0.08	-0.09	-0.08	-0.06	-0.06	-0.09	-0.09	-0.05	-0.05	-0.11	-0.07	0.09	-0.06
	Health & Lifestyle	0.16	0.11	0.18	0.16	-0.13	-0.28	-0.09	0.13	-0.31	-0.15	-0.10	-0.06	-0.07	0.10	1.00	-0.06	-0.08	-0.07	-0.05	-0.05	-0.08	-0.08	-0.02	-0.04	-0.10	-0.06	-0.08	-0.06
	Factual Knowledge & News	0.09	0.12	0.03	-0.09	-0.08	0.07	-0.10	-0.02	0.30	0.13	-0.11	-0.07	0.23	-0.08	-0.06	1.00	0.03	-0.09	-0.06	-0.06	0.09	-0.09	0.07	-0.05	-0.12	-0.07	-0.10	-0.07
	Business Profession	0.26	0.17	0.22	0.24	-0.05	-0.15	0.00	-0.04	0.16	-0.05	-0.14	-0.08	-0.04	-0.09	-0.08	0.03	1.00	0.18	-0.08	-0.08	-0.05	-0.11	-0.19	-0.06	-0.14	-0.02	0.06	-0.08
	Real Estate	-0.12	0.09	-0.01	-0.09	0.11	0.08	0.11	0.12	-0.04	0.05	-0.13	-0.08	0.19	-0.08	-0.07	-0.09	0.18	1.00	-0.07	-0.07	-0.11	-0.10	-0.18	-0.06	-0.13	-0.08	-0.11	-0.07
	Sports	-0.19	-0.10	0.07	-0.09	-0.09	0.05	0.15	-0.20	-0.11	0.08	-0.10	-0.06	-0.07	-0.06	-0.05	-0.06	-0.08	-0.07	1.00	-0.05	-0.08	-0.08	0.09	-0.04	-0.10	-0.06	-0.08	-0.06
	Games	-0.25	-0.10	0.01	-0.29	-0.02	-0.03	0.03	0.10	0.20	0.12	0.15	-0.06	-0.07	-0.06	-0.05	-0.06	-0.08	-0.07	-0.05	1.00	-0.08	-0.08	-0.13	-0.04	-0.10	-0.06	-0.08	-0.06
	Technology	-0.03	-0.20	-0.02	0.12	0.44	-0.07	0.11	-0.14	-0.07	0.17	0.02	0.14	-0.01	-0.09	-0.08	0.09	-0.05	-0.11	-0.08	-0.08	1.00	-0.11	-0.13	-0.07	-0.15	-0.02	-0.12	-0.08
	Travel	-0.03	0.09	0.41	0.15	-0.09	0.08	0.00	-0.01	0.08	-0.12	-0.13	-0.08	-0.10	-0.09	-0.08	-0.09	-0.11	-0.10	-0.08	-0.08	-0.11	1.00	-0.19	-0.06	0.61	-0.09	-0.12	0.08
	Society	-0.13	-0.05	-0.29	-0.16	-0.15	0.38	-0.21	-0.08	0.14	-0.05	-0.09	0.17	-0.16	-0.05	-0.02	0.07	-0.19	-0.18	0.09	-0.13	-0.13	-0.19	1.00	0.01	-0.14	-0.15	-0.15	-0.14
	Ethics & Philosophy	-0.08	0.27	-0.15	0.18	-0.13	0.08	0.11	0.06	-0.18	0.27	-0.08	-0.05	-0.06	-0.05	-0.04	-0.05	-0.06	-0.06	-0.04	-0.04	-0.07	-0.06	0.01	1.00	-0.08	-0.05	-0.07	-0.04
	Recreation	-0.01	0.14	0.08	0.22	0.00	0.07	-0.13	-0.03	0.07	-0.30	-0.12	-0.10	-0.13	-0.11	-0.10	-0.12	-0.14	-0.13	-0.10	-0.10	-0.15	0.61	-0.14	-0.08	1.00	0.05	0.13	-0.10
	Homes & Garden	-0.03	0.16	0.01	0.01	0.08	-0.05	0.27	-0.03	0.07	-0.12	-0.11	-0.07	-0.08	-0.07	-0.06	-0.07	-0.02	-0.08	-0.06	-0.06	-0.02	-0.09	-0.15	-0.05	0.05	1.00	0.10	-0.06
	Lifestyle	0.03	-0.11	-0.18	0.03	0.43	-0.08	-0.08	-0.02	0.13	0.03	-0.02	-0.09	-0.11	0.09	-0.08	-0.10	0.06	-0.11	-0.08	-0.08	-0.12	-0.12	-0.15	-0.07	0.13	0.10	1.00	-0.08
	Trivia	0.01	-0.14	0.07	0.04	-0.18	0.08	-0.09	0.07	-0.02	0.06	-0.10	-0.06	-0.08	-0.06	-0.06	-0.07	-0.08	-0.07	-0.06	-0.06	-0.08	0.08	-0.14	-0.04	-0.10	-0.06	-0.08	1.00

Table 8: Correlation between dimensions and content categories of information needs (Experiment 3, Pearson's r)

VARIABLE	COEFFICIENT	STD. ERROR	T VALUE	P(> |T|)
(Intercept)	0.1654	0.0742	2.23	0.0305
(Dependence on Time)	0.1651	0.1185	1.39	0.1699
Layman	0.2531	0.0778	3.25	0.0021
Operator	-0.0945	0.0605	-1.56	0.1248

Table 9: Linear regression model to explain degree of sociality, Residual standard error: 0.0733 on 48 degrees of freedom, F-statistic: 5.675 on 3 and 48 DF, p-value: 0.00, Adjusted R-squared: 0.2157

12.6.2 *Explaining Sociality*

Apart from general correlations of attributes (as discussed in the previous section), it is also interesting how SOCIALITY can be explained using the other variables as explanatory variables. A high degree of explanation could suggest that social interaction plays an important role in some specific content areas and that some attributes of information needs would encourage the use of social means to satisfy the information need (and vice versa, i.e. some information needs and content areas are not suited for social information retrieval). Therefore, a linear regression model was fitted based on the dimensions shown above. After applying an optimization using the Bayesian Information Criterion (BIC), the only factors kept are DEPENDENCE ON TIME, LAYMAN, and OPERATOR. Table 9 shows the results for the linear model. The model's residuals are distributed normally to a sufficient degree (studentized Breusch-Pagan test: p-value = 0.36, Goldfeld-Quandt test: p-value = 0.80), and the residuals are not autocorrelated (Durbin-Watson Test, p-value: 0.65). As already seen in the previous section using the correlation coefficients, the categories LAYMAN and DEPENDENCE ON TIME positively correlate with SOCIALITY – however, only LAYMAN is statistically significant ($p = 0.00$). OPERATOR has a negative impact on SOCIALITY (but the result is statistically not significant with $p = 0.12$).

When fitting and optimizing a linear regression model for the content categories (cf. Table 10), FINANCE & INSURANCE and HEALTH & LIFESTYLE have a negative impact on SOCIALITY due to negative factors in the linear model (-0.1331 and -0.1622). Both values are statistically significant ($p = 0.03$, $p = 0.01$). FACTUAL KNOWLEDGE & NEWS has a positive impact (coefficient: 0.2727) on a statistically significant level ($p = 0.01$). While a negative correlation with FINANCE & INSURANCE can intuitively be explained, given the maturity, seriousness, and high personal impact of the domain, HEALTH & LIFESTYLE and FACTUAL KNOWLEDGE & NEWS are unexpected. HEALTH & LIFESTYLE could be explained by the fact that people would like to consume passive information and have only a limited disposition to discuss individual problems with other users. The residuals of the model are normally distributed (Shapiro-Wilk normality test: W = 0.97, p-value = 0.17), studentized Breusch-Pagan test: p-value = 0.74, Goldfeld-Quandt test: p-value = 0.11) and no autocorrelation can be shown (Durbin-Watson Test, p-value = 0.20).

VARIABLE	COEFFICIENT	STD. ERROR	T VALUE	P(> |t|)
(Intercept)	0.2925	0.0125	23.40	0.0000
Entertainment	-0.0713	0.0434	-1.64	0.1076
Finance & Insurance	-0.1331	0.0580	-2.30	0.0264
Health & Lifestyles	-0.1622	0.0604	-2.69	0.0101
Factual Knowledge & News	0.2727	0.1044	2.61	0.0122
Games	0.0999	0.0605	1.65	0.1059
Ethics & Phiosophy	-0.1110	0.0653	-1.70	0.0961

Table 10: Linear regression model to explain degree of sociality using content categories, Residual standard error: 0.0711 on 45 degrees of freedom, F-statistic: 4.035 on 6 and 45 DF, p-value: 0.00, Adjusted R-squared: 0.2631

12.6.3 *Summary*

The findings suggest differences in the degree of sociality and mobility for various content areas and other attribute types. The most obvious finding is that laymen positively correlate with fields that can be characterized by a large degree of sociality. Factual Information & News (or opinions on these topics), Games, and Business Profession show the highest correlation with Sociality, while Health & Lifestyle is negatively correlated.

12.7 LIMITATIONS

The findings of the conducted experiment need to be interpreted carefully: The experiment covers only a limited sample of websites and it is not possible to guarantee that the randomly chosen URLs reflect the assigned content categories completely. In addition, the axes which were chosen to classify information are based on initial assumptions, but can not be considered exhaustive. The existence of other suitable axes is quite likely (e.g., degree of emotionality or degree of assurance).

13

EXPERIMENT 4: ROUTING OF INFORMATION NEEDS

The following experiment uses an inverted index to represent terms as vectors in a vector space defined by Wikipedia articles. This index was created as part of Oriana Baldizan's Master's Thesis (Llanes, 2016) (October 2015 to April 2016), which was supervised by Christoph Fuchs and Georg Groh at the Chair for Applied Informatics – Cooperative Systems at Technische Universität München.

13.1 SYNOPSIS

In this experiment, we compare various routing mechanisms based on topic models (LDA) and ESA with each other and traditional metrics (TF, TF-IDF) to identify expertise using a publicly available data collection with 1,400 scientific abstracts including author information, queries, and relevance judgments covering aeronautical engineering. The abstracts are interpreted as knowledge profile in a social information retrieval scenario (cf. Chapter 7). Our results suggest that both LDA and ESA can solve the routing problem, whereas the LDA-based approach and an ESA approach considering links perform best on the tested dataset.

13.2 MOTIVATION

One of the most important steps for an information seeker in social search is to identify the recipients of a query. As outlined in Section 7.1, the proposed routing mechanism is based on three components:

1. the expertise and knowledge of the potential information provider,
2. the distance in the social network, and
3. the social capital market model.

In this experiment, we focus on the first component, identifying expertise, for the following reasons:

- Tools like topic models and ESA (including Wikipedia's graph structure) offer new interesting possibilities to identify expertise.
- The effect of the social network is already covered in Experiments 1 (cf. Chapter 10) and 7 (cf. Chapter 16).
- Evaluating the social capital market model (cf. Section 7.1.3) would require a much more complex experiment design and dataset and is therefore subject of future work.

The objective of the following experiment is to relate expertise descriptions (i.e., the abstracts of each author) to semantically rich information structures (i.e., the Wikipedia concepts or the latent topics in LDA) and to use this structure to identify expertise to answer information needs (i.e., the provided queries).

13.3 RESEARCH QUESTION

The main research question for this experiment can be stated as follows:

> How do more complex routing mechanisms based on topic models (LDA) or Explicit Semantic Analysis (ESA) perform in comparison with approaches inspired by traditional information retrieval techniques like TF or TF-IDF?

A better performance would justify the additional complexity and computation effort over traditional algorithms like TF-IDF. Furthermore, a more compressed representation of expertise represented by Wikipedia concepts (ESA) or latent topics (LDA) would be beneficial when defining the knowledge profile an information provider would like to share (cf. Chapter 7).

13.4 DATASET

The dataset required for this experiment needs to contain (1) knowledge profiles for individual users, (2) a set of queries covering topics related to the knowledge profiles, and (3) for each query, users who are able to process the request (i.e., the ground truth baseline used for comparison with the results of the tested approaches). A publicly available dataset that meets all requirements mentioned above is the Cranfield collection[1]. It consists of 1,400 scientific abstracts, author information for each abstract (1,390 authors in total), and 226 queries with relevance judgments (i.e., pointers to abstracts in the collection which are relevant to the respective query). For the experiment, each abstract is interpreted as part of an individual knowledge profile of the abstract's authors. Queries are regarded as exemplary queries from an information seeker who knows the knowledge profile of each author in the dataset. The authors of the abstracts which are relevant to a query (according to the relevance judgment supplied with the data) are regarded as the correct recipients of the query (i.e., the ground truth information). For ESA, a database dump of the English Wikipedia (created on 2015-09-01) is used.

13.5 APPROACH

As a first step, author names are manually normalized to improve the quality of the profiles (e.g., to reliably identify the authors for abstracts of papers written by multiple authors). After this step, the dataset contains 1,390 authors. Then, the data is preprocessed: common stopwords, words shorter than 3 letters, and punctuation

1 http://ir.dcs.gla.ac.uk/resources/test_collections/cran/ (retrieved 2015-12-13)

are removed, text is converted to lower case, and words are stemmed with Porter's algorithm[2].

For each of the mechanisms explained below (TF, TF-IDF, LDA, ESA, ESA-IDF, ESA-LINK), the following steps are performed:

- For each query, a sorted list of authors who will most likely be able to answer the query (identified using the respective mechanism) is generated.
- The results are compared with the ground truth information and precision-recall curves are calculated (where precision and recall values are plotted for increasing result sets, i.e., the first point is defined by the precision and recall values when only the first result is received, the second point is defined by precision and recall values when only two results are received, etc.). A good algorithm can be characterized by high precision and recall values.

In the following paragraphs, the various mechanisms to calculate suitable authors for a query are presented.

TERM FREQUENCY (TF) Term frequency is used to identify relevant authors: knowledge profiles and queries are represented as vectors where each dimension corresponds to a term. The magnitude of the dimension indicates the frequency of the respective term in the document that is represented by the vector (in our case, a knowledge profile of an author or a query). Vectors are compared using cosine similarity, a well-established standard in vector space models (cf. Section 2.3.2).

TERM FREQUENCY · INVERSE DOCUMENT FREQUENCY (TF-IDF) TF-IDF enhances the TF approach explained above, the only difference is that IDF is also considered (inverse document frequency, cf. Section 2.3.2). Practically, this means that terms that occur only in a small fraction of the knowledge profiles are considered with a higher weight.

TOPIC MODELS/LDA Each author's knowledge profile consists of a topic model with ten topics, calculated using Latent Dirichlet Allocation. Each topic (and each query) is represented as a probability distribution over terms and can therefore be compared using Jensen-Shannon Distance (cf. Section 5.4). To select a profile, the following two approaches are considered:

- Maximize single topic similarity (LDA-SINGLE): select profile with the *topic that has the highest similarity to the query*
- Maximize average topic similarity (LDA-AVG): select profile where the *average similarity between the query and all other topics in the knowledge profile is maximized*

EXPLICIT SEMANTIC ANALYSIS (ESA) Using ESA, the knowledge profiles are represented using vectors in the Wikipedia concept space. As already explained in Section 5.5, each term in a knowledge profile or a query can be represented as a vector in

2 Using the implementation in the Natural Language Toolkit (NLTK), `http://www.nltk.org/` (retrieved 2015-12-16)

the Wikipedia concept space. The dimensions in this vector space represent Wikipedia articles, the value of the vector $\vec{v}$ in a specific dimension d is defined by the TF-IDF value of the respective term (which is represented by $\vec{v}$) and the Wikipedia article which is represented by dimension d. Following this approach, it is possible to represent knowledge profiles and queries as vectors (and compare them using cosine similarity). To reduce the number of dimensions (the original Wikipedia database contains more than 20 million articles), we use the pruning approach based on a sliding-window algorithm described in (Gabrilovich and Markovitch, 2009), leading to 2.1 million remaining articles.

EXPLICIT SEMANTIC ANALYSIS AND IDF (ESA-IDF) ESA-IDF is based on the ESA approach explained above. The only difference is that an additional factor similar to IDF is used to give knowledge profiles with rare information a higher weight. After calculating the profile vectors in the traditional way as explained above, a frequency vector is calculated which contains the frequency of each dimension in the collection of knowledge profile vectors. The frequency vector has a value of x in dimension d if and only if x profile vectors have a value > 0 in dimension d. In an additional step, the value of each dimension d in each profile vector is multiplied with

$$\log\left(\frac{N_A}{f_d}\right) \tag{67}$$

where N_A reflects the number of profile vectors (= the number of authors) and f_d is the number of profile vectors which have a value > 0 in dimension d (taken from the frequency vector).

EXPLICIT SEMANTIC ANALYSIS AND LINK INFORMATION (ESA-LINK) ESA-LINK enhances the plain ESA approach with the connections between Wikipedia articles. A Wikipedia article is embedded in a network of other articles which are related in terms of content. Someone who might not have explicitly stated knowledge about a specific concept might still embody a reasonable target if the person has a large amount of expertise in semantically close areas. The complete idea is explained in Section 7.1. For this experiment, we consider only outgoing links and only one single step in the graph. The major reason for this limitation is the complexity of the calculation: when calculating the similarity between a query vector and a single knowledge profile vector, for each of the positively valued dimensions of query or profile vector, the "neighborhood" of articles/dimensions has to be identified and considered in the calculation. In our (pruned) subset of Wikipedia, a typical article has on average 22 links to other pages ("outgoing links") and is referred to from 34 articles ("incoming links"). Depending on quality and popularity of the article, it is possible that an article has more than 200 outgoing or incoming links ($> 8,200$ articles have more than 200 outgoing links, $> 32,000$ articles have more than 200 incoming links). If considering the knowledge in related concept areas improves the performance, this effect should already be measurable with the limitations stated above. To compare a query

vector $\vec{q} = (q_1, q_2, ...q_n)^T$ with a knowledge profile vector $\vec{k} = (k_1, k_2, ..., k_n)^T$, the following formula is used:

$$sim(\vec{q}, \vec{k}) = \sum_{i}^{n} \left(\sum_{j \in l(i)} \frac{q_i \cdot k_j}{d(i,j)+1} \cdot \frac{1}{\sqrt{\sum_i^n q_i^2} \cdot \sqrt{\sum_i^n k_i^2}} \right) \tag{68}$$

In the definition above, $d(i,j)$ expresses the distance (length of the shortest path) between the Wikipedia articles related to dimensions i and j. $d(i,j)$ is used to reduce the positive effect of existing knowledge in related areas with increasing distance between the requested dimension and the available dimension. The function $l(i)$ defines the set of dimensions that are close to dimension i in the Wikipedia link structure, i.e. $x \in l(i) \Leftrightarrow d(i,x) \leqslant \epsilon$ for a parameter ϵ. In the experiment, ϵ is set to 1 and only outgoing links are considered.

13.6 RESULTS

The performance of the different algorithms is illustrated in Figure 10. Figure 10a shows that LDA-AVG is clearly outperforming LDA-SINGLE, TF, and TF-IDF, especially in the beginning (the most important part). Once half of the relevant items are received, all four approaches perform nearly equally well. Figure 10b suggests that ESA-LINK provides slightly better results than ESA-IDF, which in turn performs better than ESA alone. Comparing ESA-LINK, LDA-AVG, and LDA-SINGLE (cf. Figure 10c), all three perform nearly equally well (with LDA-AVG and ESA-LINK having a small edge). A detailed comparison of the differences for precision and recall of ESA-LINK and LDA-AVG is shown in Figure 9 (ESA-LINK - LDA-AVG). The line for precision shows a slight trend in the area below 0 (which would mean that LDA-AVG is the superior approach), however, the effect is quite small and more likely caused by the nature of the data than by the superiority of the LDA-AVG approach.

13.7 LIMITATIONS

The results are based on the analysis of a single collection, covering only a narrow content area (aeronautical engineering). The approximation of the users' expertise using the article abstracts does not holistically represent each user's expertise – in a real setting, expertise would be distributed heterogeneously. Furthermore, only a small subset of possible configurations was tested (e.g., 10 topics for LDA with fixed hyper-parameters 0.5, i.e. $\frac{1}{\#topics}$ and single step scenario in ESA-LINK with outgoing links only).

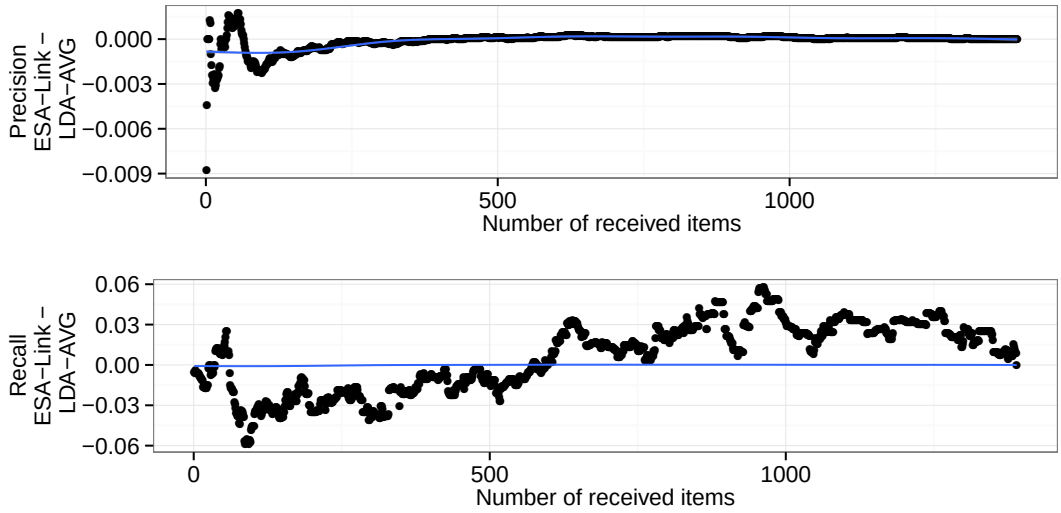
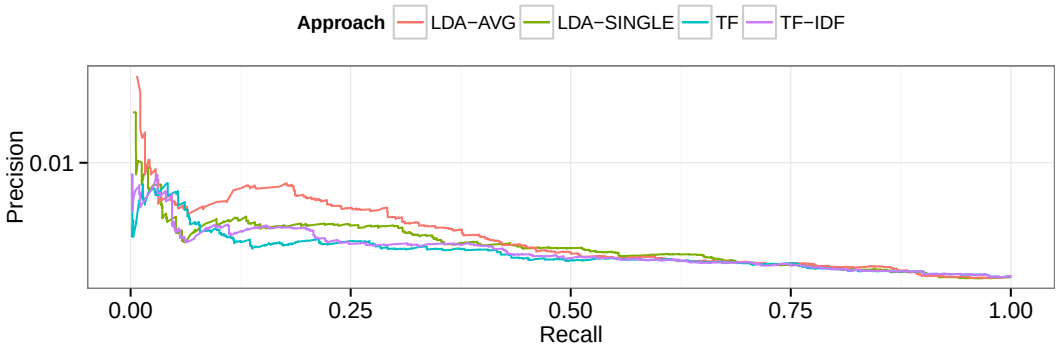
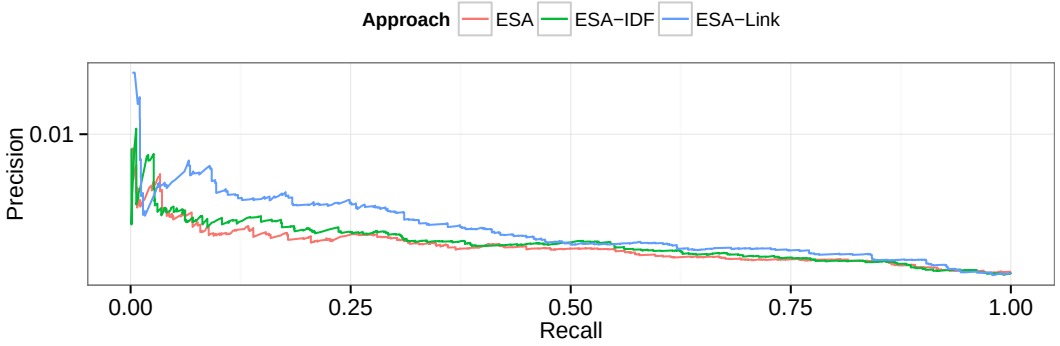


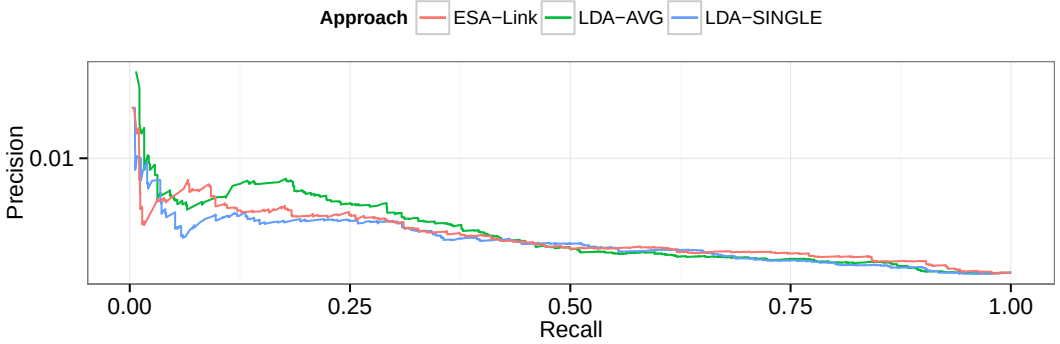
Figure 9: Difference of precision and recall values for ESA-Link and LDA (ESA-Link - LDA) on the Cranfield collection (Experiment 4)



(a) LDA, TF, and TF-IDF



(b) ESA, ESA-IDF, and ESA-Link



(c) ESA-Link, LDA

Figure 10: Performance of different routing strategies on the Cranfield collection (y-axis is log-scaled)

14

EXPERIMENT 5: INFORMATION RETRIEVAL USING TOPIC MODELS

The following experiment was conducted as part of Cordt Voigt's Master's Thesis (Voigt, 2015) which has been created between April and October 2015, supervised by Christoph Fuchs and Georg Groh at the Chair for Applied Informatics – Cooperative Systems at Technische Universität München.

14.1 SYNOPSIS

The following experiment evaluates the suitability of latent semantic spaces of documents for Information Retrieval tasks using a dataset[1] obtained from the Q&A community Stackexchange[2]. In addition, the ability of the latent semantic spaces to reconstruct human relevance judgments is explored. The latent semantic spaces are generated with Latent Dirichlet Allocation (LDA, cf. Section 5.4). In the first part of the experiment, a series of ad-hoc information retrieval tasks is performed, interpreting closeness in the latent semantic space as a criterion for relevance. In the second part, it is investigated whether the latent semantic representation allows to infer user defined quality assessments of answers. The findings suggest that the latent semantic spaces show a correlation between query and relevant information items, however, the algorithm is outperformed by a simple Vector Space Model using TF-IDF (cf. Section 2.3.2). In addition, no significant correlation between the user defined order of relevant answers to a question and the similarity-based order (using closeness in the latent semantic space as similarity function) could be demonstrated.

14.2 MOTIVATION

In a social information retrieval scenario that recommends information items to an information provider who received a query (cf. Part II), the information space model is of crucial importance for the quality of the overall system. The main purpose of this experiment is to investigate the suitability of latent semantic vector spaces of documents (based on LDA) to conduct social information retrieval tasks. The Stackexchange datasets consist of answers ("queries") and replies by the community in various topic domains. The datasets therefore have been created as part of a social process and will most likely reveal similar characteristics as any other information space in a social information retrieval scenario. LDA as mechanism to transfer natural-language text to a vector space has been chosen for the following reasons:

- Previous results (Experiment 4, cf. Chapter 13) suggest that LDA may be a valid approach to generate knowledge profiles for users representing their individual expertise.

1 `https://archive.org/details/stackexchange` (retrieved 2016-03-03)

2 `http://stackexchange.com/` (retrieved 2016-03-03)

- Latent topics allow to aggregate content in a semantically meaningful way, offering a convenient (and semantically meaningful way) to reflect privacy preferences (e.g., sharing specific topics with other users or not).

14.3 RESEARCH QUESTIONS

The experiment tries to answer the following research questions:

I. Can a latent topic representation be used to determine a relation of an answer to a question?
II. Is similarity in a latent topic space a reasonable quality measure for answers to a given question (using quality judgments of a social community as ground truth)?
III. Is text length a reasonable proxy to predict quality of answers for a given question?

Research Question I evaluates to which degree LDA can be used as an ad-hoc IR method relying on latent topic spaces. Research Question II investigates whether collaborative quality assessments of answers to questions correlate with similarity of question and respective answers in the latent topic space. If Research Question II reveals a relationship between similarity in the topic space and socially acquired quality measures, it could be possible that this relationship is caused by the fact that longer replies have a higher chance of covering more latent topics and therefore increase the probability of a higher similarity value with a question. Hence, Research Question III explores the relationship between the reply's text length and the social quality assessment done by the users.

14.4 DATASET

The basis for this experiment is a large data corpus obtained from the Stackexchange communities[3]. In recent years, Question&Answer (Q&A) communities like Quora[4], Yahoo! Answers[5], or StackOverflow[6] have become increasingly popular. The most obvious difference between Q&A sites and classical discussion boards is that the Q&A sites try to increase the quality of their content by employing several rating processes. In the Stackexchange communities, these processes include the following concepts:

- *Reputation*: Based on each user's activities in the community, each user holds a certain reputation value. Reputation can be increased by writing good answers to questions or by asking good questions. To actively influence the community (e.g., vote on answers), a certain level of reputation is required in order to prevent unexperienced users to negatively impact the community's processes.

- *Up- and Down-Votes*: Each user can give positive or negative judgments on published content (questions, answers) using the up- and down-vote mechanisms.

3 http://stackexchange.com/ (retrieved 2015-11-05)
4 https://www.quora.com/ (retrieved 2015-11-05)
5 https://answers.yahoo.com/ (retrieved 2015-11-05)
6 http://stackoverflow.com/ (retrieved 2015-11-05)

	Average size of post in characters	Average number of answers	Percentage of relevant unique terms	Number of threads	Total number of answers
Beer	101.74	2.27	21.43	**350**	**793**
German	**85.89**	**2.45**	**19.99**	4611	11,304
History	**184.91**	1.95	24.80	3792	7394
Islam	171.51	**1.79**	23.99	3575	6411
Travel	121.01	1.81	**26.14**	**10,509**	**18,991**

Table 11: Features of the selected datasets for Experiment 5; lowest and highest value for each column emphasized

This human quality assessment is used as an ordering metric to display content, e.g. the answer with the largest number of up-votes is listed right next to the respective question. Following this approach ensures that the users see the most relevant replies to a question first.

- *Accepted Answer*: The question asker can mark one of the replies as "accepted answer", which will be listed on top of the answers. This feature allows the question's author to overrule the community decision with regard to the most useful answer based on up- and down-votes.

The Stackexchange community consists of several websites covering one topic, respectively. The datasets for this experiment are taken from the Beer, German, History, Islam, and Travel communities. The datasets differ in multiple ways (cf. Table 11):

- *Size of posts*: The average length of a post varies between the communities. Posts in the History dataset are twice as long as posts in the German dataset.
- *Number of answers*: The number of answers to a question differs between the datasets and could be interpreted as an indicator for the community's activity.
- *Percentage of relevant unique terms*: As explained in the next section, words that are too rare or occur too often are removed from the dataset. The percentage of words that is considered in the analysis differs slightly between the datasets.
- *Size*: Depending on the popularity of the topic, the community and the age of the website, the five communities range from 350 threads (Beer dataset) to 10,509 threads (Travel dataset). The same logic applies to the total number of answers.
- *Writing style*: The words used and the style of writing is influenced by the discussed topics: the Beer dataset for example contains technical terms describing the brewing process, beer tasting, or stocking, while the Travel dataset consists of references to countries and popular destinations for vacation.

Apart from showing different metadata characteristics, the selected communities also have various degrees of correlation with sociality (when mapping the topics to the content categories used in Experiment 3, cf. Chapter 12, Table 7, Table 8). While Beer, German, and History can be seen as FACTUAL KNOWLEDGE (Spearman's rho: 0.23, Pearson's r: 0.30). Islam could be seen as ETHICS & PHILOSOPHY (Spearman's rho: −0.17, Pearson's r: −0.18) or SOCIETY (Spearman's rho: 0.18, Pearson's r: 0.14). For Travel, there is already a specific community (TRAVEL, Spearman's rho: 0.07, Pearson's r: 0.08). The datasets are available online in the Internet Archive[7] and can be downloaded as XML dump.

14.5 APPROACH

14.5.1 *Preparation*

Each community's XML dump is parsed and preprocessed with the following steps:

1. Import XML data to SQLite database and remove code blocks or citations within posts
2. Remove short words with less than three letters
3. Convert terms to lowercase
4. Tokenize the posts to unicode tokens
5. Stem the tokens
6. Normalize the tokens
7. Apply TF-IDF and remove all words that appear in less than five and in more than 50% of the documents to (a) focus on the most characteristic words and (b) reduce the complexity of the following operations

Afterwards, a topic model is fitted to the preprocessed data (or a subset of it) and all questions and answers are translated to their topic representation, i.e. converted to a vector in a latent topic vector space. For all answers and questions, a similarity score is calculated using the Jenson-Shannon divergence (cf. Section 5.4.2). After manually inspecting the word clouds and making sense of the results for several configurations, the number of topics for the LDA algorithm was set to 100.

The result is a SQLite database file that contains the questions and answers, the scores of the answers as given from the up- and down-votes, and different meta-information (creation date, number of answers per section, etc.). In addition, a similarity table is created which stores a similarity score for each question and each answer in the latent topic space. The topic representations and the derived similarity database depend on the part of the corpus that is used to generate the topic model. In this experiment, three different training sets for the topic models are compared: (1) the questions only, (2) the answers only, and (3) the complete corpus (i.e., the questions and the answers).

7 `https://archive.org/details/stackexchange` (retrieved 2016-03-03)

Figure 11: Word clouds for the German dataset, covering the topics "grammatical cases" and "pronunciation" (Experiment 5) (Voigt, 2015)

To calculate topic models, Blei et al.'s LDA (Blei et al., 2003) is used. The resulting latent semantic space can get illustrated using word clouds, where font size and color show the probabilities of the words that constitute the respective topic. Figure 11 shows the word clouds of two latent topics in the German dataset (limited to 100 words in the graphical representation). The first topic clearly deals with grammatical cases in German, while the second covers pronunciation-related entries.

14.5.2 *RQ I: Using Similarity in the Latent Topic Space for IR*

The performance of the following five IR approaches will be compared:

- Question TM: Latent topic space defined using *questions* to train the topic model
- Answer TM: Latent topic space defined using *answers* to train the topic model
- Question and Answer TM: Latent topic space defined using *questions and answers* to train the topic model
- TF-IDF: Vector space defined using traditional *TF-IDF metric* (cf. Section 2.3.2)
- TF: Vector space defined using *term frequency* alone

The restriction of available data to train the topic models can be seen as an analogy to an information retrieval system that only has access to a subset of the data (and therefore needs to convert between different vector spaces). The TF-IDF approach has been included as a benchmark to represent a centralized IR system. In addition, TF is used to simulate a naive system that only uses information taken from the (isolated) document itself.

Information Retrieval systems can be evaluated using Precision and Recall as performance metric (cf. Section 2.3.4). For each of the datasets and each of the five IR approaches, a precision-recall curve is calculated following these steps:

1. For each question q that has at least one answer:
   a) Get a list A^q of all answers in the dataset sorted by similarity to q
   b) Get the list with answers A^q_{rel} for q (i.e., all answers given to question q)
   c) In the following, store the precision and recall values based on looking at the first i elements in A^q in p^q_i and r^q_i respectively
2. Calculate the average p_i and r_i for all questions, i.e.

$$p_i = \sum_{q \in Q} \frac{p^q_i}{|Q|}$$

   and

$$r_i = \sum_{q \in Q} \frac{r^q_i}{|Q|}$$

   with Q being the set of questions with at least one relevant answer
3. Use p_i and r_i to plot precision and recall for $i \in [1, |A|]$ with A being the set of answers in the respective dataset

14.5.3 *RQ II: Similarity in Latent Topic Space as Quality Measure for Answers*

For each question, the respective answers are sorted by their up- and down-votes, received from members of the community. The second research question in this experiment investigates whether the ranking of answers for a specific question defined by the users' judgment can be reproduced by the similarity ranking in the latent topic space. For that purpose, the following steps are executed:

1. For each question q in the corpus that has at least one answer
   a) Calculate the similarity of q to all of its answers
   b) Order those answers according to their similarity with q
   c) Quantify the difference of this order and the ranking defined by the up- and down-votes by the community
   d) Add this value to a sum for all questions
2. Normalize the sum: divide it by the number of questions

The difference between the two rankings are quantified using Exact Match Distance, Deviation Distance, and Squared Deviation Distance (cf. (Ronald, 1998)). Given two rankings, S and T, the distance measures are defined as follows (with $S(i)$ and $T(i)$ being the element on the i-th position in the ranking):

EXACT MATCH DISTANCE Exact Match Distance penalizes differences in both rankings directly by counting the number of items that do not have the same position in both rankings:

$$d_{em}(S,T) = \sum_{i=1}^{n} x_i \text{ where } x_i = \begin{cases} 0, & \text{if } S(i) = T(i) \\ 1, & \text{otherwise.} \end{cases}$$

The calculated distance can get normalized to

$$\bar{d}_{em}(S,T) = \frac{1}{n} d_{em}.$$

This metric would assign two identical rankings that are only shifted by a single position of the maximum possible distance (n or 1 in the normalized case).

DEVIATION DISTANCE Deviation Distance takes into account the differences in the position of the same item in two rankings, i.e.

$$d_{dev}(S,T) = \sum_{k=1}^{n} |i-j| \text{ where } S(i) = T(j) = k.$$

The metric can get normalized, considering the fact that the distance metric reaches its maximum for two rankings that are inverted to each other (Ronald, 1998). This maximum is $\frac{n^2}{2}$ for rankings with an even number and $\frac{n^2-1}{2}$ for rankings with an odd number of elements. Therefore, the distance metric can get normalized as

$$\bar{d}_{dev}(S,T) = \begin{cases} \frac{n^2}{2} \cdot d_{dev}(S,T), & \text{if } n \text{ is even} \\ \frac{n^2-1}{2} \cdot d_{dev}(S,T), & \text{if } n \text{ is odd.} \end{cases}$$

SQUARED DEVIATION DISTANCE The Squared Deviation Distance is defined as

$$d_{dev}(S,T) = \sum_{k=1}^{n} (i-j)^2 \text{ where } S(i) = T(j) = k.$$

In comparison to the Deviation Distance introduced above, the Squared Deviation Distance punishes rank mismatches with a quadratic term.

14.5.4 *RQ III: Text Length as Quality Measure for Answers*

In addition to the similarity measures discussed in the previous section, this section investigates whether the length of an answer is a valid proxy for the quality estimated by the participants of the community (up- and down-votes). The following steps are executed:

1. For each question in the corpus that has at least one answer:
    a) Order the answers to the question by their length (measured in characters, longest answer first)

   b) Quantify the distance between the ranking defined by the length of the answers and the actual ranking defined by the community's up- and down-votes of the answers
   c) Sum up the distances for all questions
2. Normalize the summed up distances: divide by the number of questions

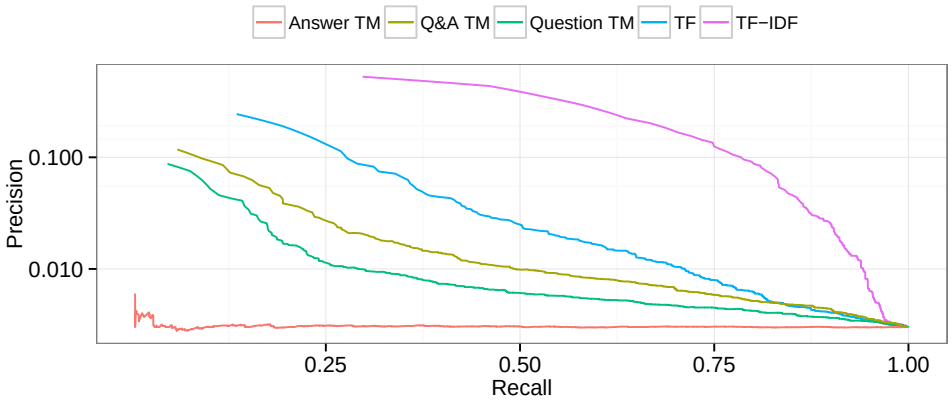
14.6 RESULTS

14.6.1 *RQ I: Using Latent Topic Space for IR*

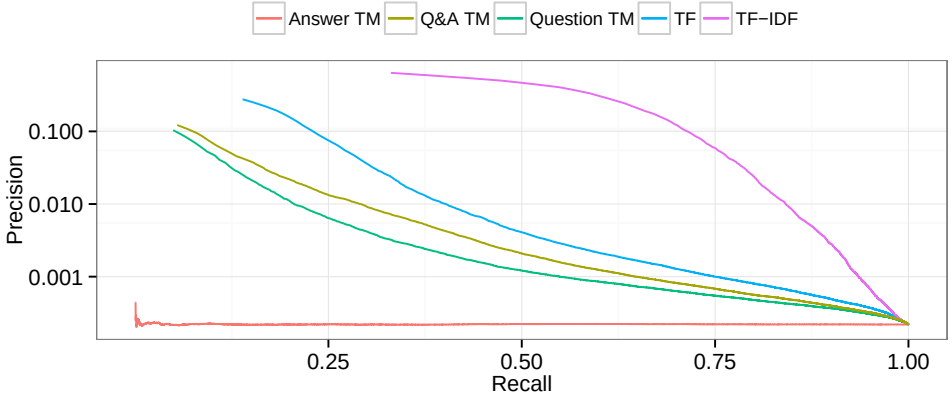
The precision-recall curves for all datasets can be seen in Figure 12 and Figure 13. TF-IDF turns out to be the clear winner when comparing it with the other approaches, especially when considering the log-scale of the y-axis. For the datasets from the Beer and German communities, TF also outperforms the topic model approaches. For the History, Islam, and Travel datasets, the approach based on the QUESTION & ANSWER TM performs best in terms of precision and recall. It is also notable that the topic model approaches QUESTION TM and QUESTION & ANSWER TM do not differ much, while ANSWER TM shows the worst performance of all approaches. This result could be caused by the fact that answers are in general shorter and do not reveal sufficient information to train a meaningful topic model. As already shown in (Cimiano et al., 2009), LDA benefits from being trained on the actual collection and not on a different background dataset. Therefore, Cimiano et al.'s findings also suggest that training the topic models on the full collection (QUESTION & ANSWER TM) increases the quality of the identified topics and their distribution. Questions seem to be better suited to reflect the collection than answers. A reason for that phenomenon could be that questions in general are formulated more detailed and cover more information than (the majority of) the answers. Another interesting finding is the relatively large performance gap between TF and TF-IDF. A possible interpretation could be that the IDF factor is an important component of TF-IDF to increase the performance of the IR system – without a measure to quantify the "scarcity" of an information item in the overall collection, finding the relevant information becomes difficult: The blue (TF) and pink (TF-IDF) lines in Figure 12 and Figure 13 demonstrate this drastically. However, calculating a metric like IDF is difficult in social settings: it would require an objective perspective on all available information items of all participating users, regardless of privacy restrictions. A potential substitute (which is not considered in this experiment) could be an individual IDF value that could be calculated for each information seeker based on the collected knowledge profiles she collected before. Following this approach, each information seeker would have an idea how scarce an information is within the published information of her social environment.

14.6.2 *RQ II: Similarity in Latent Topic Space as Quality Measure for Answers*

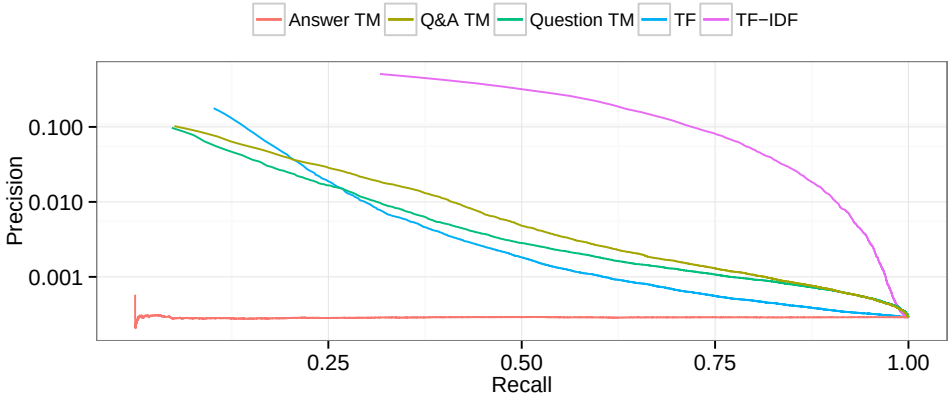
Figure 14 shows the results for the three different latent topic space approaches (topic model trained on question corpus, topic model trained on answer corpus, topic model trained on complete corpus). To correctly interpret the results, it is important to recog-



(a) Beer community

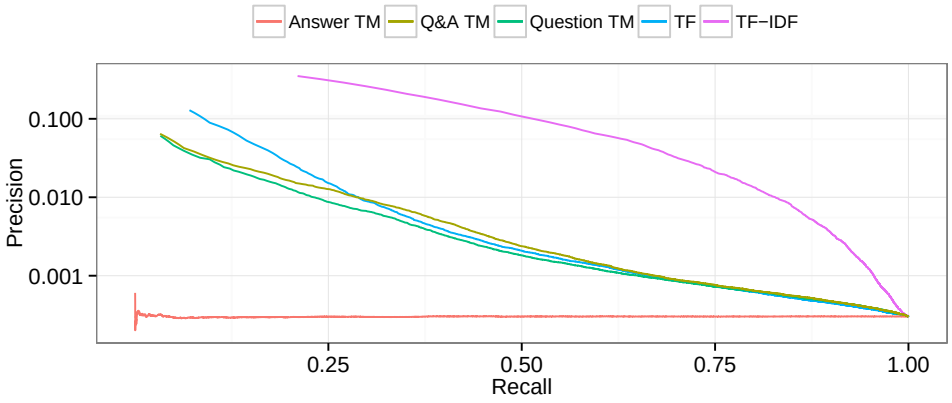


(b) German community

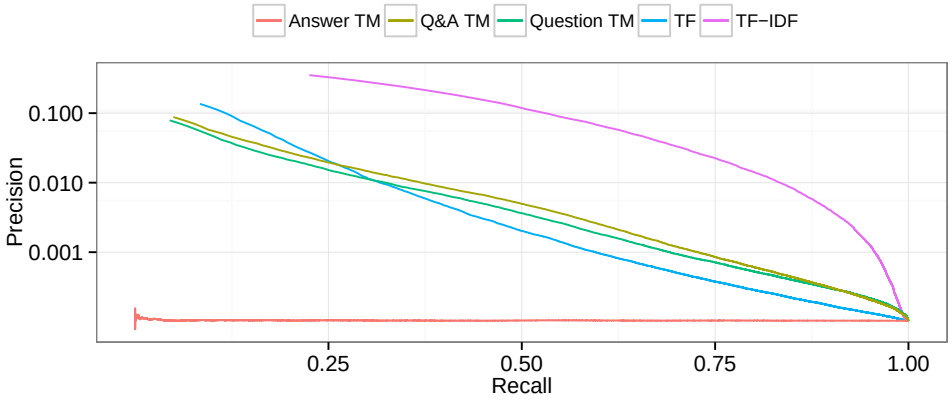


(c) History community

Figure 12: Precision-recall curves for LDA-based IR approaches and TF/TF-IDF for Stackexchange communities Beer, German, and History (Experiment 5, based on (Voigt, 2015))



(a) Islam community



(b) Travel community

Figure 13: Precision-recall curves for LDA-based IR approaches and TF/TF-IDF for Stackexchange communities Islam and Travel (Experiment 5, based on (Voigt, 2015))

nize that an identical ranking would yield a distance of 0. Due to the normalization, the largest possible distance is 1. In contrast to the previous findings in Research Question I, the different latent topic space approaches show comparable results ranging from 0.50 to 0.75. This suggests that the way of building the topic space does not seem to influence the similarity of the derived ranking to the socially collected relevance judgments (up- and down-votes). Stated differently, it can be understood as an indication that the representation in the latent topic space does not correlate with the ranking defined by the users' votes (Voigt, 2015).

To exclude the possibility that this observation is based on insufficient evidence, the experiment has been repeated, but only considering questions with at least 4 answers to get a more precise picture. The relative differences are shown in Figure 15 and are less than 10% in both directions.

14.6.3 *RQ III: Text Length as Quality Measure for Answers*

The results of the comparison between the ranking based on the length of the answers to a question and the participants' judgments (up- and down-votes) are shown in Figure 16. The same distance measures are used as in the previous section, with 0 indicating identical rankings and 1 being the highest possible distance for each distance metric. The results suggest that text length is not a suitable predicator for content quality, since the distances are located around 0.5 for all datasets (and all metrics would converge to a value around 0.5 for a random ranking, if the dataset is sufficiently large). As in the previous section, the relative difference when considering only questions with at least four answers is very small.

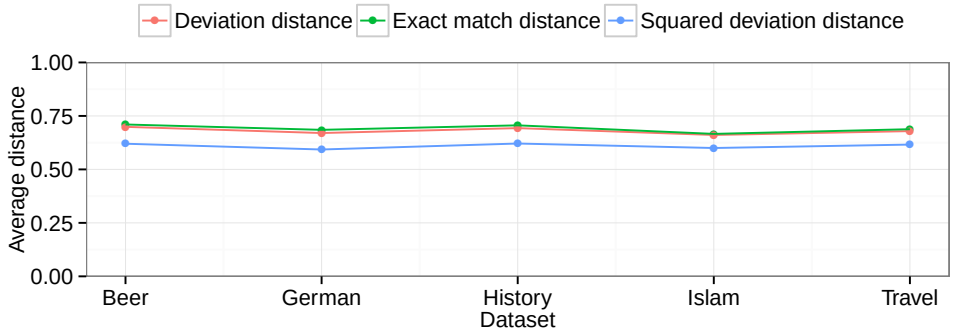
14.7 SUMMARY OF RESULTS

The findings suggest that IR approaches purely relying on latent semantic models identified via LDA perform in the same equivalence class as a plain TF approach. Given the prominent performance of the TF-IDF approach, including IDF as additional factor appears to be a promising way to yield better results. Considering the social restrictions related to IDF that have been discussed above, performing similar to plain TF might be sufficient for social applications, especially when taking into account that the topic model approach would also offer a structure that is coarse enough to define access rights and/or communicate advertising profiles (cf. Chapter 13, Section 7.1.2). Both use cases would not be possible with plain TF.

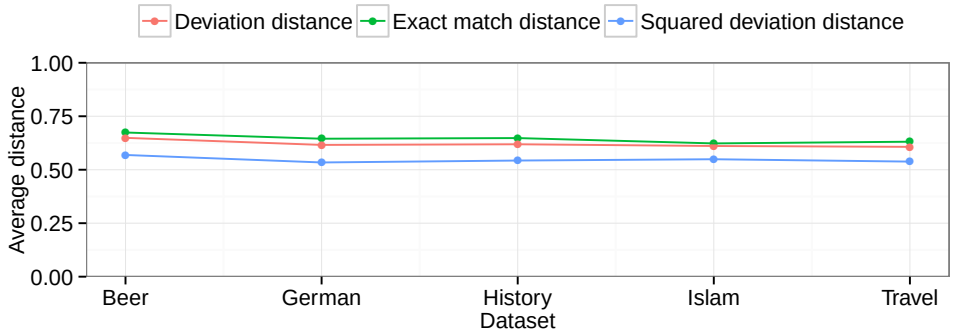
The experiments have not provided clear evidence that similarity between questions and their answers in the latent topic space defined by LDA and text length are suitable predictors for the relevance judgments assigned by human raters (i.e., members of the respective Stackexchange communities).

14.8 LIMITATIONS

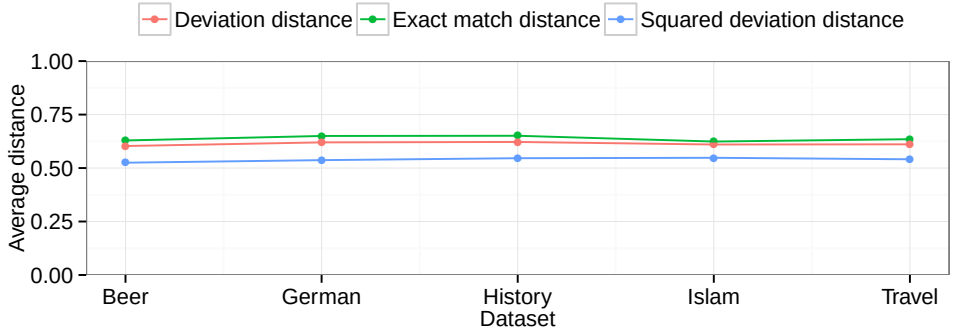
The results have to be interpreted considering a number of limitations: The number of topics for the LDA algorithm was set to 100. This was done after manually inspecting the word clouds for several configurations and trying to make sense of the



(a) Answer Topic Model

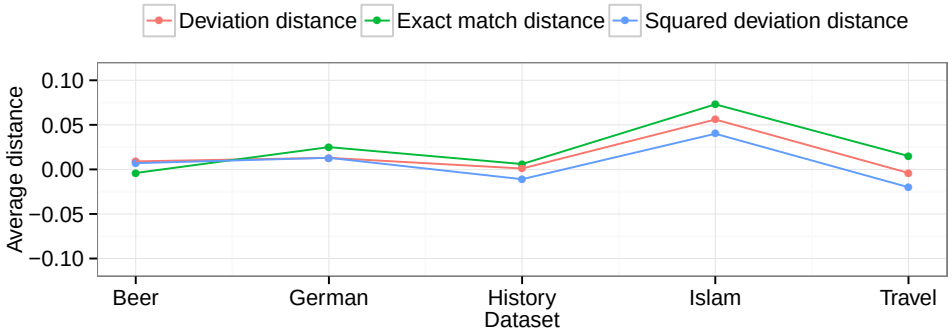


(b) Question Topic Model

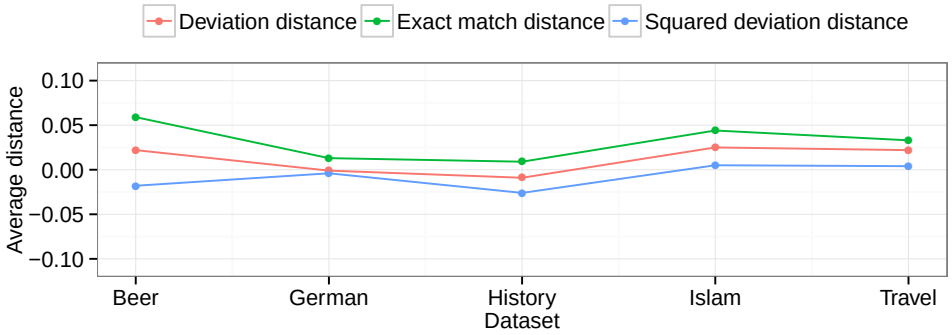


(c) Question & Answer Topic Model

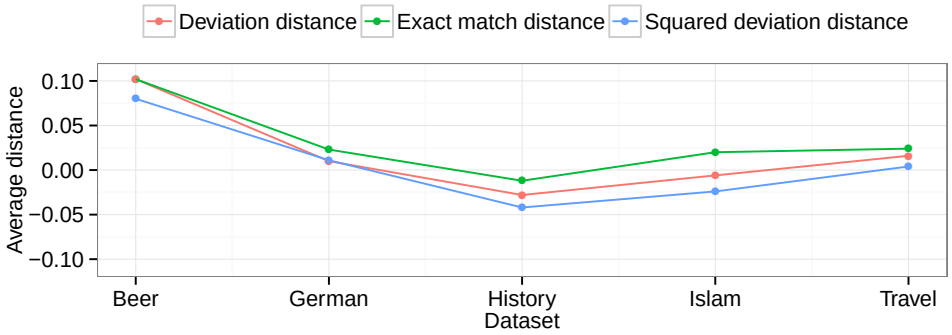
Figure 14: Average distance between ranking defined by up- and down-votes and similarity measures for questions and related answers (Experiment 5, (Voigt, 2015))



(a) Answer Topic Model

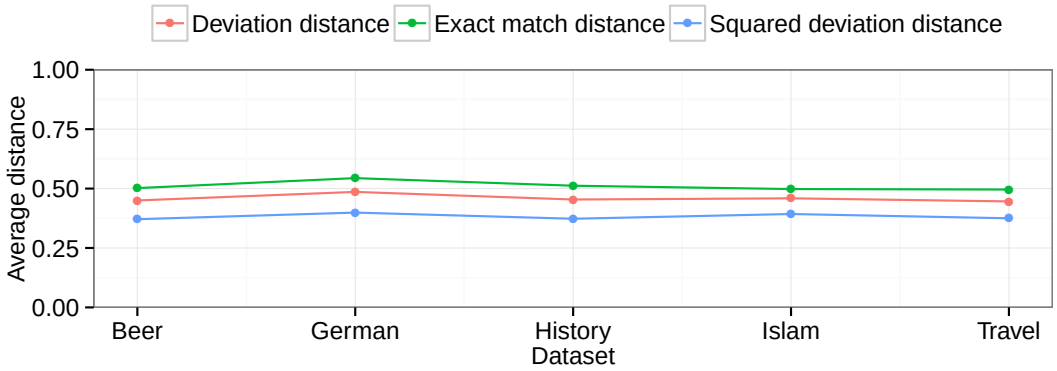


(b) Question Topic Model

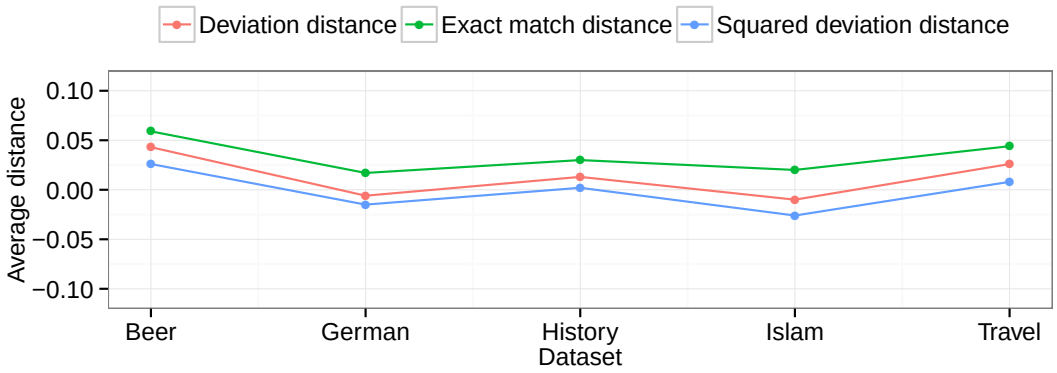


(c) Question & Answer Topic Model

Figure 15: Deviation of the results considering only questions with at least four answers (Experiment 5, (Voigt, 2015))



(a) Distance



(b) Deviation

Figure 16: Distance of rankings based on text length and community judgments (up- and down-votes); deviation of results when using only questions with at least four answers (Experiment 5, (Voigt, 2015))

results. Approaches like Hierarchical Dirichlet Processes (Wang et al., 2011) did not yield results that were easy to interpret and therefore got neglected (using the implementation in gensim[8]). In addition, adjusting the preprocessing steps (e.g. number of words to keep) might change the results and could lead to different outcomes. Due to time constraints and a slightly different focus of the thesis, we did not calculate and compare multiple configurations (e.g., the calculation of the topic models on the Travel dataset took multiple days on an Intel Core i7 CPU at 3.20 GHz with 64 GB RAM using the gensim implementation). LDA was chosen because of its popularity and high degree of maturity – using a different topic modeling approach might also lead to different findings.

In the evaluated setting, the fact that an answer has been posted as a response to a certain question defines that the answer is relevant to the question (and that all other answers which got posted as replies to different questions are not relevant). This model is a simplification of the real situation, where answers could be relevant to questions, even if the Stackexchange community dataset does not contain the link.

The whole experiment was based on the idea that all data is available, i.e. that the IR system has access to all information items. This is not the case in a distributed social information retrieval system. Therefore, the results have to be interpreted as an upper bound of precision-recall performance of the tested approaches in a situation where all information is shared and available to the IR algorithms. In particular, the IDF component, which appears to play a crucial role in the evaluated strategies requires information about the whole corpus.

8 https://radimrehurek.com/gensim/ (retrieved 2015-11-06)

15

EXPERIMENT 6: INFORMATION RETRIEVAL USING EXPLICIT SEMANTIC ANALYSIS (ESA)

The following experiment was conducted in the context of Oriana Baldizan's Master's Thesis (Llanes, 2016) (October 2015 to April 2016), which was supervised by Christoph Fuchs and Georg Groh at the Chair for Applied Informatics – Cooperative Systems at Technische Universität München.

15.1 SYNOPSIS

The experiment investigates whether expressing queries and information items in a higher-level concept space, e.g. defined by the articles of the English-language edition of the Wikipedia (ESA, cf. Section 5.5), improves the performance of information retrieval processes in comparison to traditional approaches. Using a publicly available dataset from the Stackexchange Q&A communities, we apply an IR algorithm based on ESA to identify the answers to each question in the corpus and compare the performance to traditional IR algorithms (TF, TF-IDF) and previous work (LDA, cf. Chapter 14). In comparison to the previous experiment based on LDA, where latent topics have been identified inductively, an explicit representation of concepts is used. This underlying semantical structure also allows a coarse clustering of content items in the user's information space to e.g. depict sharing preferences. Our results show that TF-IDF outperforms the ESA approach (which in turn is better than the LDA approach identified in the previous experiment, cf. Chapter 14).

15.2 MOTIVATION

The objective of this experiment is to evaluate ESA (cf. Section 5.5) in the same Information Retrieval scenario that was already used in Experiment 5, RQ I (cf. Section 14.5.2) and compare the results with the other approaches (TF, TF-IDF, LDA). The underlying hypothesis was that a "common ground" of basic knowledge where everything can get related to is helpful, especially in a case with social involvement, where IDF (cf. Section 14.7) seems to play an important role in preferring scarce features. In our experiment, the Wikipedia corpus forms the common ground, each query or information item in the information space can get interpreted as combination of multiple concepts (represented by Wikipedia articles) and get expressed as a vector in the Wikipedia article or concept space. Relying on Wikipedia adds semantically rich relations to the matching process of query and information item and could improve the IR quality. It acts as one possible common reference frame in which the effect of IDF-like model aspects are mitigated because each agent's individual clustering of information items is based on the same schema, which allows to compare structures across agents and to calculate measures similar to IDF.

15.3 RESEARCH QUESTION

As in Experiment 5, the experiment shall answer the question whether an ESA-based Information Retrieval approach can keep up with other approaches in the social IR case or even outperform them.

15.4 DATASET

The Stackexchange datasets Beer, Islam, History, and Travel already introduced in Experiment 5 are reused for this experiment. In addition, the same inverted index for Wikipedia that was already used in Experiment 4 (cf. Chapter 13) based on a Wikipedia database dump (from 2015-09-01) is reutilized (Llanes, 2016).

15.5 APPROACH

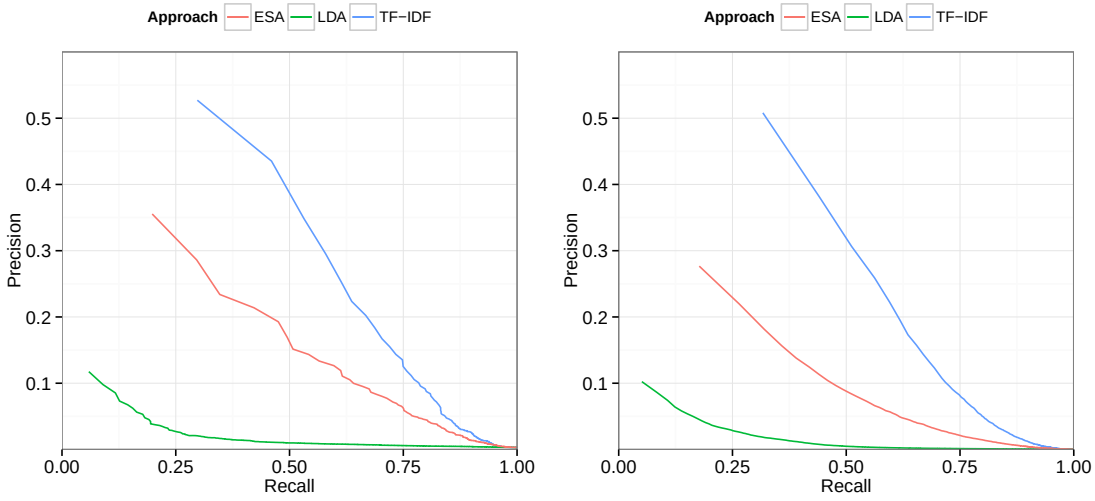
To allow a comparison of the results, a similar approach is used as in Experiment 5, RQ I, but this time we rely on ESA instead of LDA. To calculate the degree of correlation of single words with Wikipedia articles, the steps documented in (Gabrilovich and Markovitch, 2009) are followed. The resulting inverted index has also been used in Experiment 4 (cf. Chapter 13) and returns for each term the most related Wikipedia concepts (using TF-IDF). The index has been created following the steps explained in (Gabrilovich and Markovitch, 2009), also applying a pruning algorithm to reduce the index from > 4.9 million Wikipedia articles to 2.1 million articles. As in previous ESA-related experiments, vectors are compared using cosine similarity.

15.6 RESULTS

Figure 17 shows the precision-recall curves for ESA, LDA, and TF-IDF. Each point on a curve reflects an averaged precision/recall pair for a specific set of retrieved items (following the respective retrieval model for the curve), starting from the left. The first point on each curve denotes the average precision/recall combination when retrieving a single result item, the second point when retrieving two items, etc. Overall, ESA performs better than LDA on the sample datasets (from the beginning, it has better results for precision and recall), but it still does not reach the same performance level as TF-IDF. When manually analyzing example matches in detail, the results semantically often make sense (e.g., a highly relevant answer to a question actually covers an aspect of the question, but since the answer is not linked to the question in the Stackexchange corpus, it is considered as an irrelevant match, even if it might be relevant from a content perspective).

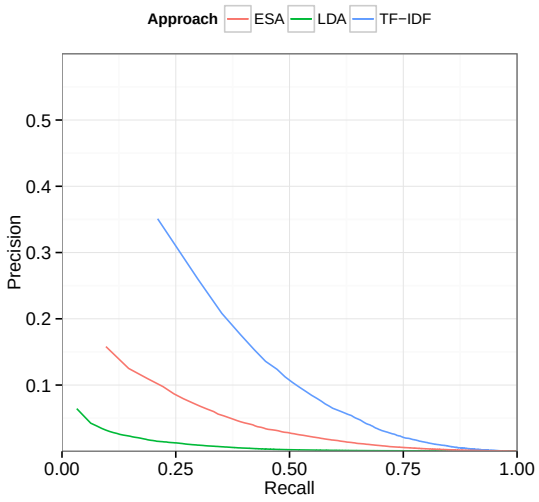
15.7 LIMITATIONS

The results must be interpreted carefully: each corpus that was used to test the various algorithms covered a very homogenous topic area. As already mentioned as limitation of Experiment 5 (Section 14.8), the fact that an answer has been posted as

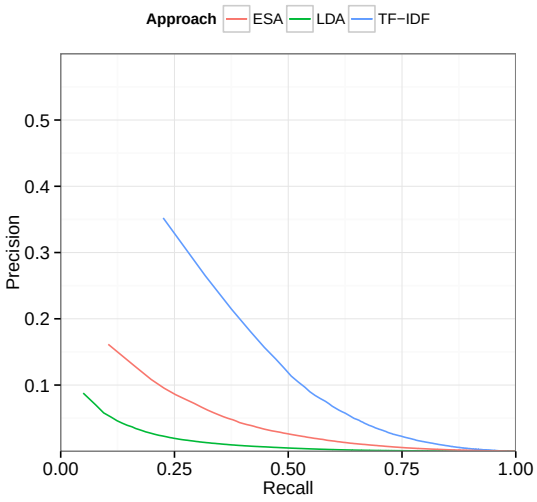


(a) Beer

(b) History



(c) Islam



(d) Travel

Figure 17: IR performance of ESA, LDA, and TF-IDF on the Stackexchange corpus (Experiment 6)

a response to a certain question does not necessarily mean that this answer is irrelevant for all other questions. However, to avoid the tedious process of identifying relevant answers for each question, this criterion has been selected as ground truth. Especially in the ESA case, a manual inspection of the similarity table revealed that similar question/answer pairs make sense from a content perspective, although they are not related to each other in the Stackexchange community. In future work, a more diverse corpus with more comprehensive relevance judgments could possibly simulate a more realistic IR scenario, which might allow LDA and ESA to demonstrate their strengths.

16

(MAIN) EXPERIMENT 7: SOCIAL INFORMATION RETRIEVAL EXPERIMENT

While each of the previous experiments covered only a small aspect of a social information retrieval scenario, the following experiment was designed to model the complete process using a prototypical implementation for multiple social information retrieval scenarios.

16.1 SYNPOSIS

We simulated several social information retrieval scenarios with 121 participants, covering (1) a manual mode (where participants can query manually selected social contacts, who reply manually), (2) an automatic mode (where the participants' queries automatically get assigned to other users by the system based on eight different routing strategies and the responses are automatically extracted from the information providers' individual information spaces), and (3) a specific use case where searching for a product in an online shop is supported by buying/viewing decisions of the participants' social environment. The experiment covers different social information retrieval scenarios, including multiple interaction types and various forms of accessing the private information spaces as discussed in Chapter 8 to be able to investigate limits and chances of social information retrieval.

Our findings suggest that information providers answer to queries with a higher probability if information seeker and provider are socially close in terms of tie strength and sympathy. The information provider's openness to share private information depends on the audience and interaction archetype: reactive sharing (i.e., actively requesting information from an information provider) appears to be a more successful pattern for social information retrieval than dealing with the more complicated generalized privacy considerations associated with relying on proactively published material.

Information seekers' relevance judgments on received replies are positively influenced by high levels of tie strength and sympathy. Serendipity appears to be positively impacted by high levels of content knowledge similarity.

When choosing designated information providers, information seekers rely on acquaintances with high values for tie strength, sympathy, social context similarity, and content knowledge similarity. Certain content (Media, Tech, Food, Travel) and query types (Recommendation, Opinion, Factual Knowledge, Personal Experience) appear to be more suitable for social search than others. In general, the new findings roughly confirm and broaden the results obtained from the previous experiments (Experiment 2 in Chapter 11 for query types, Experiment 3 in Chapter 12 for content types).

16.2 MOTIVATION

While the other experiments focused on isolated components of the social computing concept (cf. Part II), this overarching experiment involves all parts of the user's social search journey from the identification of suitable information needs, the selection of the appropriate audience (and possible alternatives), to the evaluation of response quality and impact on social relationships. For organizational reasons, the experiment has been split in three parts which have been run sequentially in May and June 2015 at TU München:

- *Experiment 7a* (Section 16.5.2) covers the scenario where users ask other (consciously selected) people for help. The addressee of the query replies manually using unstructured text (without being allowed to refer to traditional search engines to search for the answer on the web). The scenario in this experiment is also referred to as "manual mode".
- *Experiment 7b* (Section 16.5.3) is different from Experiment 7a due to the fact that the audience of the query is not chosen by the information seeker but automatically selected by a routing strategy. In addition, replies to the queries are not manually entered textual responses written by the information provider, but automatically identified URLs of websites which have been visited by the information provider in the past and which might be relevant to the information seeker's query. This approach is also referred to as "automatic mode".
- *Experiment 7c* (Section 16.5.4) depicts a scenario where the information seeker searches for a product to buy in an online store. For each product, the information seeker gets informed whether it has been bought or seen by friends. In addition to the products that match the information seeker's query, the result list of the online store also contains items that have been seen or bought by the information seeker's friends and are from the same product category the information seeker is searching in. This method generalizes the content-induced constraints for the mentioned products in the result list and creates a broader result set with more diverse items, which still might be interesting for the information seeker (same product category) and offer a chance for serendipity (caused by social closeness to the person who bought/looked at the product).

16.3 RESEARCH QUESTIONS

The social computing experiment was designed to holistically cover the intended social search scenario. Therefore, the research questions are similar to the ones mentioned earlier in Chapter 1:

1. How do (a) social context and (b) interaction archetypes influence users' data sharing sensitivity in view of social information retrieval approaches?
2. Relevance and Serendipity of Results
   a) How relevant are information items taken from non-public information spaces of socially close people when satisfying information needs?

   b) Does social context imply a valuable contribution to retrieving information from the unconscious information need (serendipitous information)?

3. Which social concepts influence the users' routing decisions?
4. Which categories of information needs could benefit from social information retrieval?

16.4 PARTICIPANTS

The participants of all three subexperiments have been students of the *Social Computing* and *Social Gaming* courses held during summer term 2015 at TU München. The experiments have been part of the students' voluntary homework assignments, which consisted of various exercises of the social network analysis realm based on the collected (and anonymized) data. The students could improve their overall grade of the respective course by submitting the homework assignments. For students who did not want to participate in the experiment, an alternative set of exercises was prepared so that they would not face any disadvantages. The students have been informed about the research questions and the procedure of the experiment in advance without revealing significant design decisions (e.g., that we added artificial recipients in experiment 7b who automatically replied with results taken from a traditional search engine). The participants where asked to act as if they were using the real system; no incentive was given to intentionally influence the result of the experiment (e.g., it was clearly stated that the number of recipients a student adds to a query does not impact the evaluation of her homework). During the experiments, students had the possibility to use a web forum to clarify questions on homework assignments and report technical problems.

16.5 APPROACH

16.5.1 *Preparation*

The experiment is designed as a within subject study, i.e. each participant joined all three subexperiments (7a-7c). During the preparation phase, the participants registered in a web interface, imported their social network exported from Facebook, and assessed their social contacts (i.e., evaluated tie strength, sympathy, social context similarity, and content knowledge similarity for all other participants of the experiment and 50 random friends out of the Facebook contact list to reduce the workload, Table 12). To protect the participants' privacy, only hashed names got transmitted. Table 37 details the variables obtained during the preparation phase. The technical architecture and an overview of the developed tools is also given in the appendix (cf. Section A.2.2). The participants received a comprehensive introduction: a series of screencasts was created to explain each of the steps in the experiment, the related homework assignment, and the required tools. In addition, a written documentation was provided. In case of any questions/problems, the participants could post their concerns in a web forum or contact the author via email, phone, or in person.

Variable	Scale type	Question
Tie strength	ratio	"How strong is your relationship?" (0: weak, 100: strong, default: 50)
Sympathy	ratio	"How much sympathy do you have for the other person? (your input will not be disclosed to anyone)" (0: not likeable at all, 100: highly likeable, default: 50)
Social context similarity	ratio	"How similar are your social contexts?" (0: very different, 100: very similar, default: 50), incl. link to get a definition of social context ("Social context is defined by the social setting you are living in: Sharing the same workplace, school, course, friends, etc. with a friend would imply similar social contexts").
Content knowledge similarity	ratio	"How similar is this person to you in terms of content knowledge?" (0: very different, 100: very similar, default: 50, optional checkbox: "I do not know whether this person's content knowledge is similar to mine or not")

Table 12: Variables describing the relationship between participants and their social network imported from Facebook in Experiment 7

In case the user did not have a Facebook account (or assumes that this account does not provide any useful information because it is not actively used), we asked her to import social contacts derived from her email inbox using the following process to get a list of friends:

- Sort email inbox alphabetically by sender
- Add the first 50 names of human senders (i.e., no mails automatically generated by IT systems) to the list of social contacts

Only 2 participants (out of 121 who submitted their social network) had to use this alternative approach.

16.5.2 *Experiment 7a: Social Information Retrieval in Manual Mode*

In the first part of the experiment, participants were asked to route (1) three self-defined queries and (2) three predefined queries to potential information providers in their individual social network. Each participant had to enter the self-defined queries before getting to know the predefined ones to allow an unbiased definition of information needs. The information providers were contacted via email by the information seekers and received a custom link to the web system to reply to the query or to forward it to someone else (either inside or outside the set of experiment participants) who might be better suited to give a response. In either case, the information provider filled out an online survey to allow us to understand the reasons for her decision and give further details on the response (level of privacy, whether it has already been shared before, information about the relation to the information seeker, etc. – please refer to Table 38 for a complete list). The information provider was asked to only use already known information in the reply, i.e. she should not start a search session on the web herself to find an answer (however, searching to rediscover something was

allowed). After the information seeker received the information provider's response through the web system, the information seeker was also asked to fill out a survey to evaluate the reply (satisfaction with response, relevance, degree of personalization, whether the reply contained unexpected information, information about the relation to the information provider, etc. – please refer to Table 38 for a complete list). To protect the anonymity of the information providers, no personal information that could help to identify them (email address, name, etc.) was asked by or saved on the web system. The complete social network on the server was based on the hashed version generated by SNExtractor (see Section A.2.2) – the participant's computer was the only place where the real names of the information providers have been stored.

The three predefined queries have been selected based on (Oeldorf-Hirsch et al., 2014) and (Fuchs and Groh, 2015b) to ensure that we have a minimum set of queries which are known to perform well in social information retrieval scenarios:

1. "Where can I grab some cheap and good food near [TU Munich] Stammgelaende?"
2. "Which movie would you recommend to watch with a close friend?"
3. "My parents will visit me next week - do you know any good place to eat?"

Table 38 lists the variables obtained in the manual mode of the social information retrieval experiment from information seeker and information provider. In addition, Table 39 lists the additional variables used when the original information providers (referred to as IP^n) decides to forward the query to someone else (referred to as IP^{n+1}).

The system ensures that the social tie for each selected information provider has been assessed by the information seeker (if the assessment did not take place using SNExtractor, it has been done during the setup of the manual query).

16.5.3 *Experiment 7b: Automated Social Information Retrieval Using Topic Models*

In the second part of the experiment, participants built a representation of their information space, asked queries, and replied to them based on potentially relevant items in their information space. Therefore, the group of participants was limited to the registered participants (in contrast to Experiment 7a, where it was possible for participants to route queries to Facebook contacts outside of the registered group of participants).

The experiment's model of a participant's information space consisted of the content of the websites that have been read by the participant. The participant's browsing history represented the source for URLs, which was downloaded and was used to build a custom topic model for each participant. This process took place on the participant's local computer using the tools described in Section A.2.2 – no browsing histories were uploaded to the central server for privacy reasons. Participants had the possibility to remove items from their browsing history during the process. For a real-world social information retrieval application, we would assume that a suitable privacy function exists (cf. Section 8.3) that defines whether an information item is allowed to be shared with an information seeker, given a specific query. In the absence

of such a function, the decision is done manually by the user to ensure that the user's privacy concerns are met.

Even with TF-IDF having performed better in classic IR scenarios (cf. Chapter 14, Chapter 15), we opt for topic models for the following reasons:

- Granularity is inferred inductively, i.e. topic categories are calculated based on the user's private information space and therefore offer a classification that is semantically closer to the user's content than other approaches (e.g., ESA, which is based on Wikipedia and therefore introduces a potential distortion due to the standardized concepts).
- Topic models offer a more flexible structure than plain term-based approaches like TF/TF-IDF – relying on terms directly for the matching process would make an intuitive privacy configuration based on semantically meaningful chunks practically impossible due to the number of dimensions and their missing semantic relation.

Using the browsing history as basis for the information space was motivated by three reasons:

- The data is available – the browsing history is accessible, HTML or PDF documents can get downloaded and included in the topic models easily. More sophisticated approaches which might include other types of data (like oral communication) would require much more effort.
- The data represents information the participant consumed. Depending on the individual preferences, a significant amount of information is gained through the web. Given the fact that all participants are students and enrolled in a course related to computer science it is likely that their information gathering behavior is dominated by the internet and therefore a significant proportion of their recently acquired knowledge is covered.
- The data also documents decisions the user made, like items that got bought from online stores or documents that got read. This transactional type of data forms the most valuable component in the user's information space because it reveals the results of the user's complex opinion-forming process and also reflects a significant part of the user's consumed knowledge (which might have been processed further to form new artifacts).

Once the information spaces have been generated locally on the participants' computers and the derived topic models have been uploaded to the web system, the participants initiated queries using the web system. In contrast to Experiment 7a, the participants could only specify the query, but not its recipients. The recipients were chosen by the system with one of the following eight strategies:

- *Tie strength*: Select those participants as recipients of the query initiated by the information seeker IS who have a very high (low) tie strength with IS (assessed by IS).

- *Sympathy*: Select those participants as recipients of the query initiated by the information seeker IS who are considered as very (not) likeable (at all) by the information seeker IS.
- *Social context similarity*: Select those participants as recipients of the query initiated by the information seeker IS who have a very high (low) similarity of social context with IS (assessed by IS).
- *Content knowledge similarity*: Select those participants as recipients of the query initiated by the information seeker IS who have a very high (low) similarity of content knowledge with IS (assessed by IS).

For each initiated query, one of the eight routing strategies was selected randomly to define a set of three designated information providers. Using the uploaded topic models representing the information providers' information space, relevant URLs were identified in each information provider's information space using Jensen-Shannon divergence (cf. Section 5.4.2). Each information provider was notified by the web system that some of the URLs she visited are considered relevant for a received query (we chose to return the five URLs which were ranked best). The information providers received the IDs of the respective URLs that have been considered relevant and were asked to translate each ID back to the URL using the URLTranslator tool to ensure privacy (cf. Section A.2.2). In a production-ready social information retrieval system, this step would have been automated on the user's device: for the experiment, we decided to minimize the complexity of the applications which are required to run on the user's device to avoid the additional challenges by supporting and testing multiple platforms. Therefore, the matching process to identify relevant information items in the information provider's information space for a given query has been executed on a central server (instead of a local device which could access the user's complete information space directly). To still maintain the information provider's privacy, we only asked the user to upload her trained topic models (and not her complete information space) to the central server. Since the uploaded topic models do not allow to infer the related URL of the information item (as it is kept on the user's device) in this prototypical implementation, the matching process needs the information provider's support to map the information item ID to a real URL string.

The information providers could decide to stay anonymous (i.e., the information seeker would not know who replied to the query), but they had to provide a reason for that (open question).

In parallel, the server sent each query also to a traditional search engine (BING[1]) in the background and created a fake response containing the first five results for the information seeker. In 50% of the cases, the fake accounts returned their results anonymously. The fake replies were visible to the information seeker with a random time delay (up to 12 hours) to simulate the behavior of a normal human participant. This kind of reply was given to allow a comparison of replies obtained from social means and traditional search engines. BING was chosen as a search engine because of the possibility to query it using an API[2].

1 `http://www.bing.com` (retrieved 2015-05-01)

2 BING Search API, `https://datamarket.azure.com/dataset/bing/search` (retrieved 2015-05-01)

Once a new reply was available for the information seeker, she got informed and was asked to evaluate the response quality using the variables listed in Table 13.

Variable	Owner	Question
Relevance of URL	IS	"Is this link relevant to answer your query?" - scale from 0 ("Not relevant at all") to 100 ("Highly relevant") using slider, default: 50
Unexpectedness of URL	IS	"Did the website contain information you did not expect or that was not obvious; did the content surprise you?" - scale from 0 ("Content did not surprise me") to 100 ("Content was highly unexpected") using slider, default: 50
Degree of personalization	IS	"Was the link you received personalized to you as a specific person? (e.g. did the information provider include personal knowledge about you in the reply)" - scale from 0 ("Not personalized at all") to 100 ("Highly personalized") using slider, default: 50
Communal sharing	IS	"Do you think that the person who answered your query would help you no matter what, i.e. in any situation?" - scale from 0 ("No") to 100 ("Yes") using slider, default: 50
Authority rank	IS	"To which degree do you think that the person helps you because of differences in social rank or status (e.g. boss vs. staff, caring parent vs. child)?" - scale from 0 ("Low") to 100 ("High") using slider, default: 50
Equality matching / market pricing	IS	"Do you think that you owe something to the person who answered your query because she/he did you a favor by answering your question?" - scale from 0 ("No") to 100 ("Yes") using slider, default: 50
Satisfaction with results	IS	"How much did the information provided by this person help to satisfy your information need?" - scale from 0 ("Didn't help anything") to 100 ("Information need fully satisfied") using slider, default: 50

Table 13: Variables in Experiment 7b (automatic mode)

16.5.4 *Experiment 7c: Social Product Search*

In Experiment 7c, social information retrieval was exemplified using a specific use case: product search. The participants were asked to upload the products they viewed and bought from a large online retailer (Amazon[3]) to the web system. The participants had to search for ten products they would consider buying in a special user interface in the web system. The web system received the original results for the search query from the retailer's website and marked those items that have been bought or viewed by a social focus group. The strategies to identify the focus group match the routing strategies explained in Section 16.5.3; for each query, one of the eight strategies was chosen randomly. In addition, products from the same product category

3 `http://www.amazon.com` (retrieved 2015-10-17)

that were bought or viewed by the focus group and which have a small distance to the query, evaluated by a fuzzy string matching function, were added to the result set. The information seeking participant was asked to rate the usefulness and the degree of unexpectedness of each result item using a 5-star rating interface. In addition, clicks on the items got recorded.

Using this approach, we wanted to investigate whether socially close people can offer information that is regarded relevant by the participant (Research Questions 2a and 2b, cf. Section 16.3). This would support a broader view on relevance (cf. Section 2.3.5) which might not only be based on matching certain algorithmic criteria relying on the content of the item, but also on the social closeness of people.

Table 14 lists the variables that were collected during the experiment.

Variable	Owner	Question
Usefulness of item to query	IS	"Please rate usefulness (on 5-star scale)"
Unexpectedness of item to query	IS	"Please rate degree of unexpectedness (on 5-star scale)"
Click	IS	(User clicked on product item)

Table 14: Variables in Experiment 7c (product search scenario)

16.6 RESULTS

16.6.1 *Experiment 7a: Social Information Retrieval in Manual Mode*

The collected data from Experiment 7a provides relevant results for all four research questions mentioned in Section 16.3. In the following, the results for each research question are presented in detail.

Research Question 1 is split into two parts, the first one covers the influence of the social context (Section 16.6.1.1) and the second one discusses the results for different interaction archetypes (Section 16.6.1.1).

16.6.1.1 *RQ 1a: How Does Social Context Influence Users' Data Sharing Sensitivity?*

One of the questions we would like to answer with this experiment is whether social closeness between information seeker and provider changes the participants' data sharing behavior in such a way that more private information items are exchanged.

As explained before, the information seekers sent six queries to potential information providers. Three out of those six queries were predefined (and were the same for each participant), the remaining ones could be defined by the information seekers individually. Before sending the queries, the information seekers assessed their relationship to the information providers (tie strength, sympathy, similarity of social context, and similarity of content knowledge; please refer to Section 16.5.2 for further details). The information providers received the queries via email and could reply to the request by clicking on a link and filling out an online form. In addition to providing a response to the query, the information providers also assessed their social relationship to the information seeker and gave an estimate of the privacy level of the provided information.

INFORMATION PROVIDER'S WILLINGNESS TO REPLY TO REQUESTS Out of 925 sent requests, 21 did not contain valid estimates from information seekers for the social attributes (tie strength, sympathy, social context similarity, content knowledge similarity) due to technical problems, resulting in 904 requests with a valid estimation of social attributes by the information seekers (after specifying a query, information seekers had to assess their social relationship to the information provider if this has not already been done in the SNExtractor tool – due to a bug in the web application, the query itself was sent, even if the window where the social attributes should be entered was closed later). Out of those 904 requests, 217 did not receive an answer from the intended information provider. Figure 18 shows the distribution of tie strength, sympathy, social context similarity, and content knowledge similarity for both categories (estimated by the information seeker, 39 requests have been not considered for the analysis of content knowledge similarity, because the information seekers indicated that they can not assess the content knowledge similarity with the information provider). The Wilcoxon rank sum test confirms for sympathy and tie strength that both populations have a different mean (sympathy: W = 64,634, p-value = 0.00, tie strength: W = 63,726, p-value = 0.00). For content knowledge similarity and social context similarity, H_0 (true location shift is equal to 0) can not be rejected (content knowledge similarity: W = 69,250, p-value = 0.33, social context similarity: W = 69,898, p-value = 0.17). When fitting a logistic regression model to predict whether an information provider would reply to a request based on the estimates for tie strength, sympathy, social context similarity, and content knowledge similarity (all estimated by the information seeker, Table 43), the optimized model (Table 44) reveals a positive statistically significant impact of tie strength: *for each single increase* in tie strength, the odds to get a reply are multiplied by $\exp(0.009) = 1.009$, i.e. they are slightly increasing (one single unit increase on the tie strength scale would lead to to roughly 1% increase of the odds to get a reply). For an information seeker, it therefore makes sense to choose contacts with high tie strength to increase the probability of receiving an answer.

In the following, the information provider's perspective is explained. After removing incomplete requests (e.g., requests that did not receive an answer from the information provider or miss social attribute assessments from the information provider, especially for content knowledge similarity), the dataset consists of 605 query/response pairs. This dataset is used to evaluate the hypothesis that information providers tend to share more private information with socially close information seekers.

CORRELATION COEFFICIENTS FOR PRIVACY AND SOCIAL ATTRIBUTES Spearman's rank correlation coefficient and Pearson's product-momentum correlation coefficient are used to analyze whether the information provider's assessment of the required degree of privacy for the shared information item is correlated with the social attributes tie strength, sympathy, similarity of social context, and similarity of content knowledge. The only statistically relevant result ($p < 0.05$) is a negative correlation of privacy with sympathy (Spearman's rho: -0.105, Pearson's r: -0.100; see Table 15 for complete results).

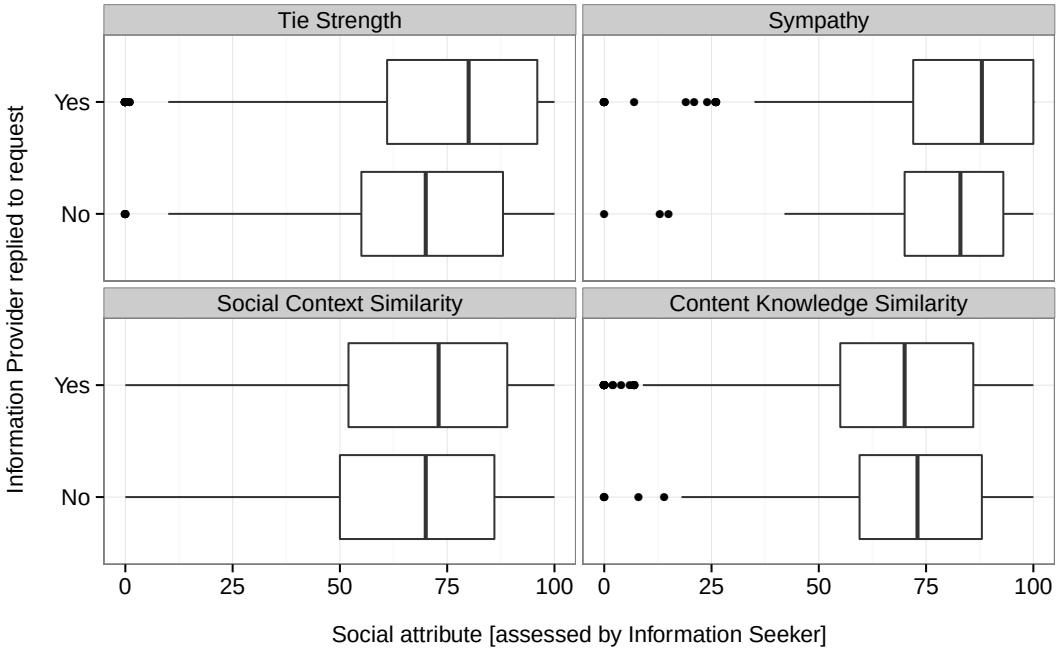


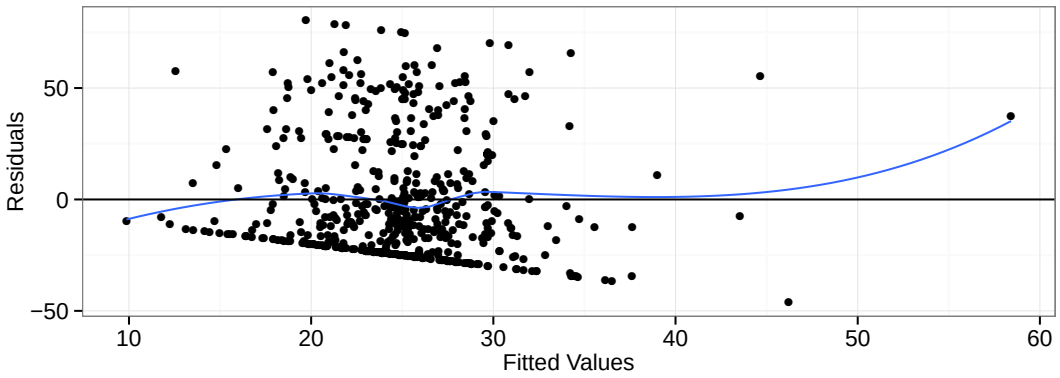
Figure 18: Ignored/answered queries (Experiment 7a)

ATTRIBUTE	SPEARMAN'S RHO	PEARSON'S R
tie strength (IP)	-0.004 (p=0.93)	0.017 (p=0.69)
sympathy (IP)	-0.105 (p=0.01)	-0.100 (p=0.01)
social context similarity (IP)	-0.058 (p=0.15)	-0.043 (p=0.29)
content knowledge similarity (IP)	-0.016 (p=0.69)	0.022 (p=0.59)

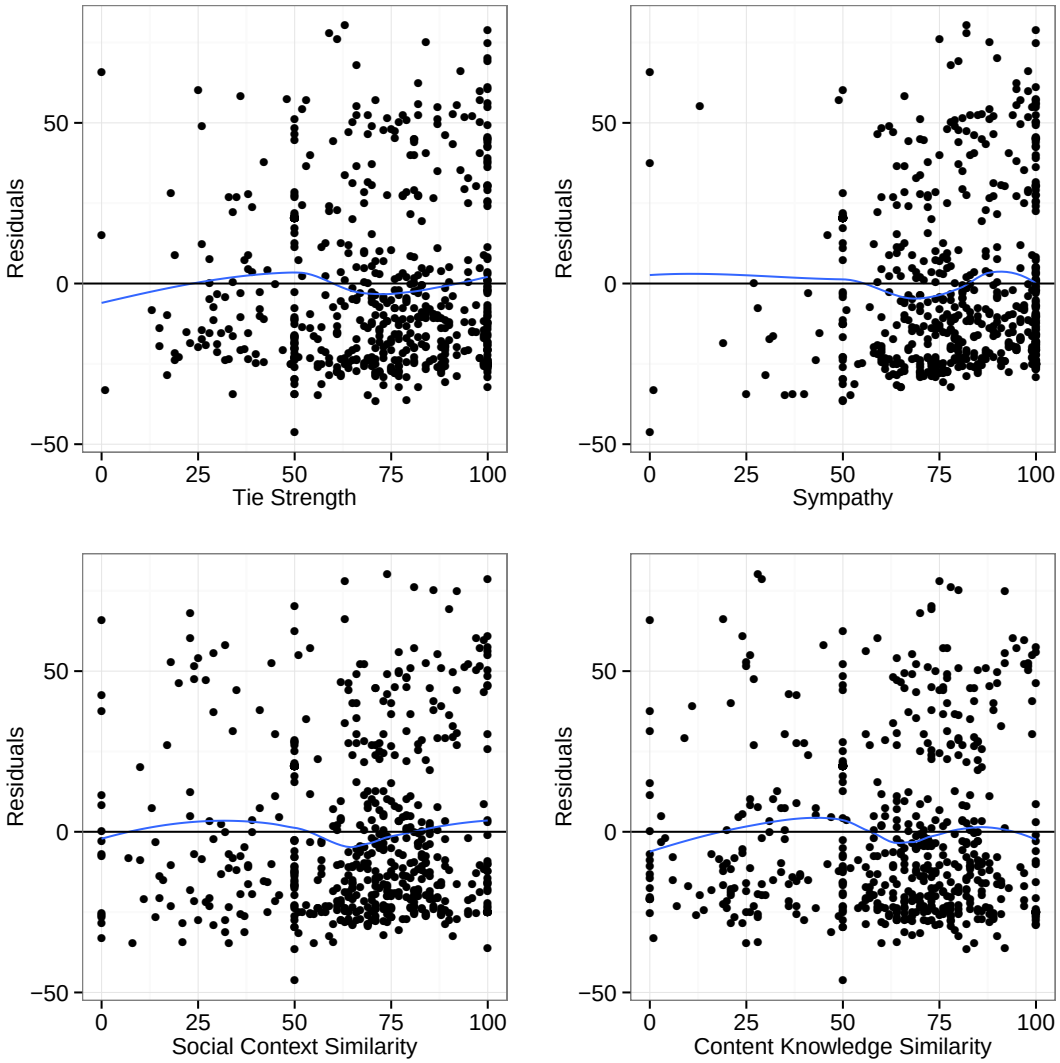
Table 15: Correlation between required degree of privacy for the shared information item and the social attributes of the relationship, assessed by the information provider, p-values based on t-distribution (Experiment 7a)

VARIABLE	COEFFICIENT	STD. ERROR	T VALUE	P(>|T|)
(Intercept)	34.2589	5.0281	6.81	0.0000
tie strength (IP)	0.2416	0.0730	3.31	0.0010
sympathy (IP)	-0.3304	0.0841	-3.93	0.0001
content knowledge sim. (IP)	0.0543	0.0492	1.10	0.2703
social context sim. (IP)	-0.0568	0.0563	-1.01	0.3136

Table 16: Properties of linear regression model to explain privacy, Residual standard error: 26.62 on 600 degrees of freedom, Multiple R-squared: 0.03003, Adjusted R-squared: 0.02356, F-statistic: 4.644 on 4 and 600 DF, p-value: 0.00 (Experiment 7a)



(a) Residuals against fitted values



(b) Residuals against predictor variables

Figure 19: Residuals of simple linear model to explain privacy (Table 16, Experiment 7a)

LINEAR REGRESSION MODEL TO EXPLAIN PRIVACY WITH SOCIAL ATTRIBUTES The coefficients and details of a linear regression model explaining *privacy* with the four social attributes are presented in Table 16. The model explains roughly 2% of the variance in the data (adjusted R^2, p-value: 0.00). Figure 19 shows the residuals for the fitted values (Figure 19a) and each predictor variable (Figure 19b). The first plot in Figure 19a reveals a certain skewness of the residuals, caused by the fact that the privacy values are limited to the interval $[0, 100]$ (this explains the horizontal boundaries in the scatterplot, the effect is much stronger for negative residuals due to the overall distribution of privacy, cf. Figure 22a). The plots showing residuals and predictor variables (Figure 19b) are quite balanced and confirm other characteristics of the data (e.g., unbalanced distribution of tie strength and sympathy). The residuals are not distributed normally (Shapiro-Wilk normality test is significant with W = 0.89714, $p = 0.00$). A more complicated and optimized (BIC, and adjusted R^2) linear model (including interactions and quadratic terms of explanatory variables) is able to explain roughly 4% of the variance in the data and is statistically significant ($p = 0.00$). Table 41 shows the explanatory variables, their coefficients, and properties. However, the model is not fully satisfactory, since the explained variance is quite low and one of the requirements of linear regression models is not met: the residuals are not distributed normally (Shapiro-Wilk normality test is significant with $p = 0.00$). Using the function $\log(x + 1)$ to transform the output variable does slightly improve the model (adjusted R^2 value is 0.05), but the residuals are still not normally distributed (Shapiro-Wilk normality test, $p = 0.00$) and the model therefore is of at least questionable expressive power (cf. Table 42).

LINEAR REGRESSION MODEL WITH RANDOM EFFECTS Each information provider introduced a certain bias in the data by her individual rating preference. In addition, different queries also impact the privacy estimation. One way of reflecting such factors is to introduce individual intercepts for each information provider and each query. This can be done using random effect models (Section 5.2.4). A linear model explaining *privacy* with the social attributes (assessed by the information provider) with individual intercepts for each information provider and each query is shown in Table 17. As expected, the residual standard error is lower than in the model without random effects (Table 16, 17.93 vs. 26.62). The coefficients are of similar size and significance as the ones in the previous model. The plot showing residuals against fitted values (Figure 20) also shows a similar pattern caused by the boundaries of the privacy variable. The residuals are not distributed normally (Shapiro-Wilk normality test, W = 0.9297, p-value = 0.00).

LOGISTIC REGRESSION MODEL TO EXPLAIN PRIVACY WITH SOCIAL ATTRIBUTES Logistic regression models do not require normally distributed residuals and therefore appear to be a suitable framework for the available data. As detailed in Section 5.2.2, logistic regression models rely on a dichotomous response variable. To transform privacy (ranging from 0 to 100) to a binary factor, 50 has been selected as threshold to classify privacy in "high" (privacy > 50) and "low" (privacy $<= 50$). Since 50 has also been the default value, the information provider had to actively move the slider to the right side (and indicate high privacy) to make the response

VARIABLE	COEFFICIENT	STD. ERROR	χ^2	P(>χ^2)
(Intercept)	35.1880	5.4539	–	–
tie strength (IP)	0.1792	0.0767	5.4619	0.0194
sympathy (IP)	-0.2691	0.0859	9.8126	0.0017
content knowledge sim. (IP)	0.0630	0.0527	1.4313	0.2316
social context sim. (IP)	-0.068	0.0595	1.2973	0.2547

Table 17: Properties of linear regression model with random effects (information provider, query) to explain privacy, p-values based on type II Wald χ^2 tests; Residual standard error: 17.93, Null deviance: 5680.2, Residual deviance: 5667.8 (χ^2: 12.412, df: 4, P(>χ^2): 0.01)

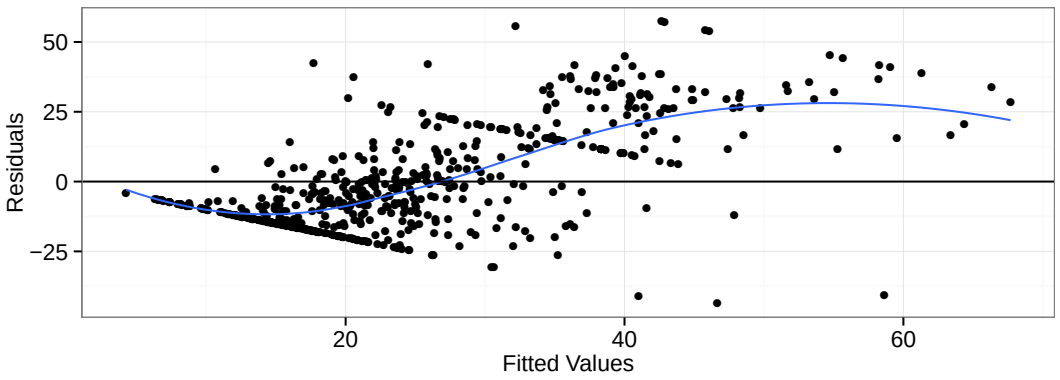


Figure 20: Residuals of linear model with random effects to explain privacy (Table 17, Experiment 7a)

VARIABLE	COEFFICIENT	STD. ERROR	Z VALUE	P(>|Z|)
(Intercept)	-3.5280	0.5999	-5.88	0.0000
tie strength (IP)	0.0217	0.0085	2.54	0.0110
sympathy (IP)	-0.0031	0.0094	-0.33	0.7431
social context sim. (IP)	-0.0031	0.0059	-0.53	0.5950
content knowledge sim. (IP)	0.0103	0.0054	1.91	0.0567

(a) Properties of logistic regression model to explain privacy_high variable (=$privacy > 50$), Null deviance: 519.22 on 604 degrees of freedom, Residual deviance: 501.62 on 600 degrees of freedom (Experiment 7a)

VARIABLE	COEFFICIENT	STD. ERROR	Z VALUE	P(>|Z|)
(Intercept)	-3.6740	0.5292	-6.94	0.0000
tie strength (IP)	0.0186	0.0059	3.15	0.0016
content knowledge sim. (IP)	0.0089	0.0049	1.84	0.0656

(b) Properties of optimized logistic regression model to explain privacy_high variable (=$privacy > 50$), Null deviance: 519.22 on 604 degrees of freedom, Residual deviance: 502.06 on 602 degrees of freedom (Experiment 7a)

Table 18: Logistic regression models to explain privacy_high variable (Experiment 7a)

part of the "high privacy" category. Out of 605 observations, 512 fall into the category "low privacy" and the remaining 93 are considered as "high privacy". Starting point for the fitted model was a model with all four social parameters (tie strength, sympathy, social context similarity, content knowledge similarity, Table 18a). Optimizing the model for AIC resulted in the model as described in Table 18b. The coefficients are logarithmic odds, so the coefficient for tie strength (0.0186) can be interpreted in the following way: for each unit increase in tie strength, the odds to fall into the "high privacy" category of the response variable get multiplied by $\exp(0.0186) = 1.0188$. A comparison of the optimized model's deviance with the null model (i.e. the plain intercept without any parameters) reveals that the residual deviance is lower (519.22 vs. 502.06, 2 df, P(>Chi)=0.0002).

LOGISTIC REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN PRIVACY WITH SOCIAL ATTRIBUTES As previously done for the linear model, it also seems to be a useful approach to account biases caused by redundant information providers or queries. A logistic regression model with random effects does not converge for all social attributes (thus, does not allow to draw conclusions from the model). Therefore, four models have been created, each containing only one of the social attributes as explanatory variable and considering information provider and query as random effects. While the model for sympathy failed to converge and the models for social context similarity and content knowledge similarity do not have significant coefficients, the model for tie strength reveals a positive coefficient for tie strength (0.0930, std. Error: 0.0213, z value: 4.365, p-value: 0.00; Null deviance: 402.44, Residual deviance: 395.56, χ^2: 6.8823, p-value: 0.01).

VARIABLE	COEFFICIENT	STD. ERROR	Z VALUE	P(>|Z|)
sympathy (IP)	-0.0257	0.0062	-4.141	0.0000
tie strength (IP)	0.0146	0.0054	2.718	0.0066

(a) Properties of ordinal logistic regression model to explain privacy on a 5-level ordinal scale

LEVEL	ESTIMATE	STD. ERROR	Z VALUE
[-0.1,20] | (20,40]	-0.7723	0.3304	-2.337
(20,40] | (40,60]	0.0041	0.3261	0.013
(40,60] | (60,80]	0.9643	0.3322	2.903
(60,80] | (80,100]	2.3060	0.3753	6.144

(b) Thresholds for privacy categories

Table 19: Properties of ordinal logistic regression model to explain privacy on a 5-level ordinal scale, as logit (Experiment 7a)

ORDINAL LOGISTIC REGRESSION MODEL TO EXPLAIN PRIVACY WITH SOCIAL ATTRIBUTES Ordinal Logistic Regression is a generalized form of logistic regression, where the response variable is not on a binary but an ordinal scale. The detailed approach is explained in Section 5.2.3. The privacy variable is transformed to an ordinal scale (with intervals [0,20], (20,40], (40,60], (60,80], (80,100]). The base model used all four social parameters as explanatory variables; sympathy and tie strength remain the only predictors for the privacy category after model optimization. Table 19 lists the coefficients and the thresholds for the different privacy categories. Like in the logistic regression model explained above, the coefficients and estimates need to be transformed with $\exp()$ to plain odds. The results can get interpreted as follows: one step increase in sympathy changes the odds for the next higher privacy category by a factor $\exp(-0.0257) = 0.9747$ (when tie strength is controlled). That means that increasing sympathy reduces the predicted probability for a higher privacy category. Analogously, one step increase in tie strength changes the odds for the next higher privacy category by a factor $\exp(0.0146) = 1.0147$, i.e. increases the predicted probability for a higher privacy category. Comparing the model with the null model confirms that the extended model (with sympathy and tie strength) fits better (Likelihood ratio tests of cumulative link models, $P(>\chi^2)$=0.00).

ORDINAL LOGISTIC REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN PRIVACY WITH SOCIAL ATTRIBUTES An ordinal logistic regression model to explain privacy with all four social attributes, considering the information provider as random effect (in the intercept) confirms the findings of the ordinal logistic regression model presented above: tie strength has a positive impact (coef.: 0.0151, std. error: 0.0060, z value: 2.540, p-value: 0.01), while sympathy negatively influences the privacy level (coef: -0.0259, std. error: 0.0068, z value: -3.797, p-value: 0.00). Social context similarity and content knowledge similarity are statistically not significant. The model's improved performance is statistically significant when compared to the null model (p-value: 0.00).

VARIABLE	TIE STRENGTH	SYMPATHY	SOCIAL CONTEXT	CONTENT KNOWLEDGE
Correlation coefficients	×	↓	×	×
Linear regression	↑	↓	×	×
Linear regr. w/ random effects	↑	↓	×	×
Logistic regression	↑	×	×	×
Logistic regr. w/ random effects	↑	×	×	×
Ordinal logistic regression	↑	↓	×	×
Ordinal log. regr. w/ random effects	↑	↓	×	×

Table 20: Summary: Impact of social attributes tie strength, sympathy, social context similarity, and content knowledge similarity (assessed by information provider) on the information provider's privacy judgment; ↑ (↓) indicates a significant positive (negative) impact, × indicates no correlation (Experiment 7a)

In the remaining part of this section, the results will be summarized for each social attribute. Table 20 provides a brief overview.

TIE STRENGTH Correlation coefficients suggest no significant correlation: values for Spearman's rho and Pearson's r are near 0; in addition, the p-values are very high (0.93, 0.69) which does not allow to reject H_0 (i.e., tie strength and privacy are not linked). In the both linear regression models (with and without random effects), tie strength has a positive significant effect on privacy. In the logistic regression model without random effects, tie strength has a slight positive effect on privacy (0.0217, p=0.01). The logistic regression model with random effects for information provider and query and tie strength as single explanatory variable confirms a positive effect of tie strength for privacy (0.0930, i.e. for each increase in tie strength, the odds for the higher privacy category get multiplied by $\exp(0.0930) = 1.0975$). In the ordinal logistic regression model, tie strength is positively linked to higher privacy levels, each step increase in tie strength increases the odds for the higher privacy category by factor $\exp(0.0148) = 1.0149$. The findings are confirmed by the ordinal logistic regression model accounting for multiple measurements (information providers) considered as random effects. Intuitively (and also confirmed by previous research, e.g. (Levin and Cross, 2004)), one would expect a positive effect of tie strength and the privacy of the information that is shared using the respective tie. Our data confirms this effect: the effect identified in the logistic and ordinal regression models is quite small (a single unit increase of tie strength increases the odds for the higher privacy category by roughly 1%) and it is not linear. It is possible that the way of measuring the variables might have influenced the results: the participants assessed privacy manually, which might have caused a bias towards extreme positions. Furthermore, the information providers only assessed their own replies, which have been triggered by the received questions. Therefore, it is possible that the queries influenced the measured privacy to a certain extent.

SYMPATHY Referring to Spearman's rho and Pearson's r, sympathy appears to be negatively correlated ($\mathrm{rho} = -0.105$, $\mathrm{r} = -0.100$) at statistically significant levels

($p < 0.05$). In both linear regression models, sympathy has a statistically significant negative impact on privacy (without random effects: -0.3304, with random effects: -0.2691). In the logistic regression models, sympathy is either not significant or the model does not converge (random effects model). In the ordinal logistic regression model, sympathy has a negative coefficient (-0.0257). One step increase in sympathy would multiply the odds for the higher privacy category with $\exp(-0.0257) = 0.9747$, i.e. decrease the odds. The findings are confirmed by the ordinal logistic regression model accounting for multiple measurements (information providers) considered as random effects. The finding is rather surprising: Intuitively, one could expect that higher levels of sympathy would foster information sharing and therefore also increase the probability for more private content – however, our data shows the opposite effect. Given the way the data was collected, it is possible that the sympathy judgments were not based on a deep relationship (like tie strength), but more on short-term effects like individual physical attraction or random encounters before or after the lecture. This could explain that even a high rating for sympathy does not necessarily lead to the required intimacy to share material with higher privacy constraints.

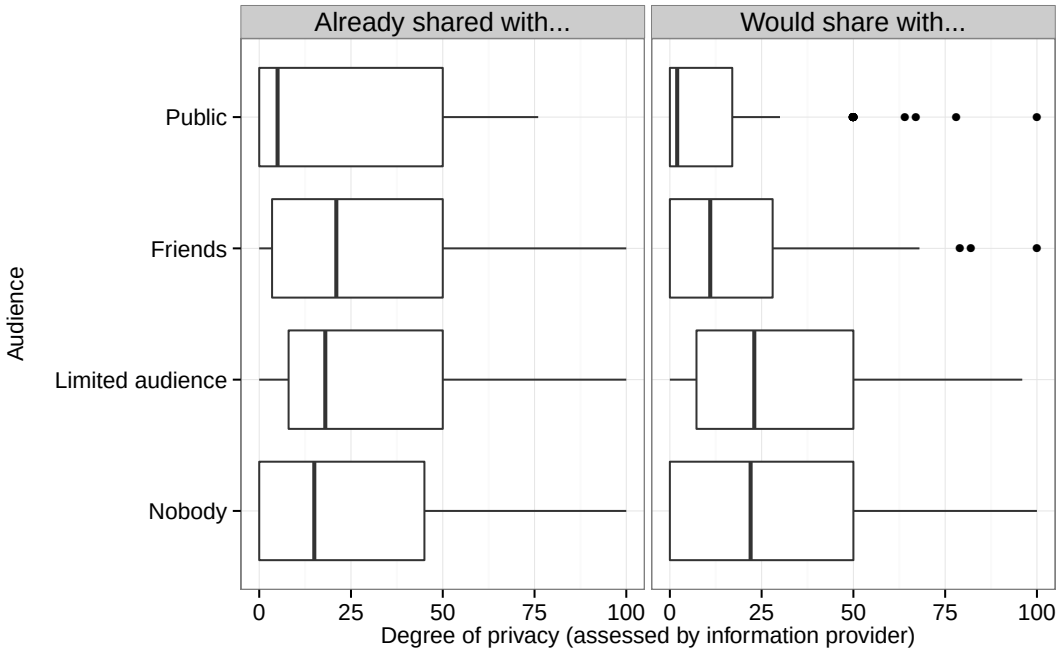
SOCIAL CONTEXT SIMILARITY The correlation coefficients do not provide any statistically significant insights for social context similarity (Spearman's $\mathrm{rho} = -0.058$, $p = 0.15$; Pearson's $r = -0.043$, $p = 0.29$). In all tested linear regression and logistic regression models, social context similarity is not statistically significant.

CONTENT KNOWLEDGE SIMILARITY The correlation coefficients do not provide any statistically significant insights for content knowledge similarity (Spearman's $\mathrm{rho} = -0.016$, $p = 0.69$; Pearson's $r = 0.022$, $p = 0.59$). In both linear and both logistic regression models, content knowledge similarity is not statistically significant.

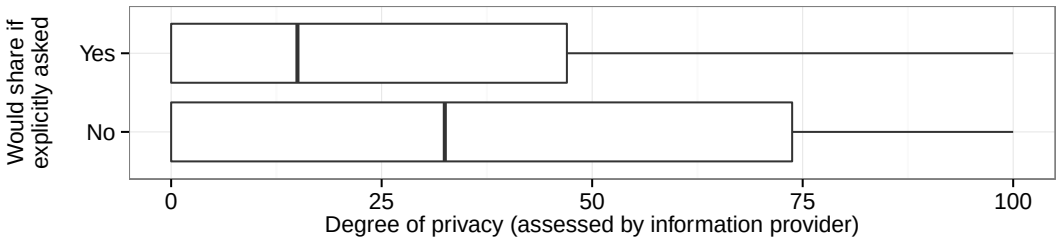
16.6.1.2 *RQ 1b: How Does the Interaction Archetype Influence Users' Data Sharing Sensitivity?*

In addition to the degree of privacy, information providers also stated whether the information they shared as a response to the query has already been shared with other users on a social network platform (available categories: No, Limited Audience, Friends, Public), whether they would share it with other users (available categories: No, Limited Audience, Friends, Public), and whether they would share it with other users if they would get explicitly asked for it (Yes, No). The analyzed dataset consists of 697 query/answer pairs, which contain the required information from the information providers. Table 21 shows the answer frequency for each category. It is remarkable that 76% of the information items have not been shared yet, despite the fact that for more than 50% of the information items the information providers could imagine sharing it with a limited audience (15%), friends (22%) or the public (16%). A majority of 95% of the information items would get shared by their respective information providers if they would get explicitly asked for it.

In the following, the association between the reported degree of privacy, the audience of an information sharing act, and the mode of the interaction (reactive vs. proactive) is investigated.



(a) Privacy degrees for information items shared (left) / hypothetically shared (right) with the respective groups



(b) Privacy degrees for information shared/not shared if explicitly asked

Figure 21: Privacy degrees of information items (hypothetically) shared in various interaction settings (Experiment 7a)

share with...	already_shared	would_share
Nobody	532 (76%)	333 (48%)
Limited Audience	72 (10%)	102 (15%)
Friends	71 (10%)	153 (22%)
Public	22 (3%)	109 (16%)
Total	697 (100%)	697 (100%)
would share if explicitly asked...		
No		32 (5%)
Yes		665 (95%)
Total		697 (100%)

Table 21: Answer frequency for different categories of sharing alternatives in Experiment 7a; differences to 100% due to rounding errors

median values Figure 21 shows the distribution of the privacy estimates for each sharing category. The plot of the information items which actually were shared (Figure 21a, left) suggests that there is no real difference between the categories Friends and Limited Audience. In fact, it seems a bit counter-intuitive that the median privacy value in the Friends group is higher than for Limited audience because the latter is less public (especially when considering the advancing trend of growing friend lists on social network platforms). As expected, information with a high degree of privacy (> 76) has not been shared with the public. When asking the information providers with whom they could imagine to share the information item on a social network platform, the plot (cf. Figure 21a, right) resembles the expected form much more: The Public category has the lowest median privacy value, followed by Friends and the more restricted category Limited Audience. The distribution of the latter does not differ much from the one of the Nobody category.

While it is not possible to reject the H_0 hypothesis that the median values for each category in the "already shared" dataset are the same using a Wilcoxon rank sum test with pairwise comparisons, the same test confirms significant differences for all categories of the "would share" dataset, except for the two categories Nobody and Limited Audience (as already expected by visually analyzing the plot).

As already stated above, when explicitly asked, the majority would share the information with the information seeker. When analyzing the median values for both categories (cf. Figure 21b), one can recognize that there are more information items with higher degree of privacy in the No category (as expected). The Wilcoxon rank sum test with continuity correction confirms that both populations do not have the same median value ($W = 13098.5$, p-value $= 0.02$).

ordinal logistic regression model to explain sharing category An ordinal logistic regression model to explain the sharing category using the degree of privacy as explanatory variable is not statistically significant for the already shared

information items ($p = 0.11$). Adding "information provider" or "query" as random effects to account for repeated measurements does not cause privacy to become significant. A model calculated for hypothetical sharing ("would share") is statistically significant (Likelihood ratio tests of cumulative link models, $P(>\chi^2)=0.00$) and confirms that an increasing level of privacy reduces the odds for a more public audience. The audience categories NOBODY, LIMITED AUDIENCE, FRIENDS, and PUBLIC are ordered as listed by increasing degree of publicness, the coefficient for privacy (-0.018) can be interpreted as multiplicative change of the odds for a more public category of $\exp(-0.018) = 0.9822$ for each increase in privacy (std. error: 0.0029, z value: -6.269, p=0.00). Adding random effects for "information provider" or "query" reinforces this effect: the model with "information provider" as random effect suggests -0.0255 as coefficient for privacy (std. error: 0.0041, z value: -6.288, p-value: 0.00), the model with "query" as random effect proposes -0.0190 as coefficient (std. error: 0.001181, z value: -16.12, p-value: 0.00).[4]

LOGISTIC REGRESSION MODEL TO EXPLAIN REACTIVE SHARING A logistic regression model for the reactive sharing scenario where the information seeker is actively asked to share information (Figure 21b) proposes -0.0195 as coefficient for privacy (std. error: 0.0059, z value: -3.282 , p-value: 0.00). Considering multiple measurements for each information provider with a random effects model (intercept only) leads to a stronger effect (coefficient: -0.0383, std. error: 0.0149, z value: -2.568, p-value: 0.01). Consequently, a single increase in privacy causes a multiplicative change of $\exp(-0.0383) = 0.9624$ for the odds to get into the YES category.

LOESS CURVES In the last analysis for this research question, for each privacy level ($0 - 100$), the ratio how many information items of the respective privacy level have already been shared/would be shared with each of the audience categories (NOBODY, LIMITED AUDIENCE, FRIENDS, PUBLIC) is calculated. For a degree of privacy p (reflected on the x-axis in Figure 22b), the position of the red point related to p shows the share of information items with privacy level p, that have been shared with NOBODY (this applies analogously for LIMITED AUDIENCE (yellow), FRIENDS (blue), and PUBLIC (green)). LOESS local regression (cf. Section 5.2.5) is used to visualize trends in the data. Figure 22b shows the ratios for *already shared* information items. The straight, almost fully horizontal lines suggest that privacy did not influence the sharing behavior of the information providers: the majority of all information items have not been shared with anyone (NOBODY), only a very small subset was shared with LIMITED AUDIENCE and FRIENDS, while near to nothing was shared with PUBLIC. This suggests that the degree of privacy did not cause any differences in the dominant sharing preference. Furthermore, in most cases the participants did not share information items, what can be seen as a limiting factor for social information retrieval approaches which rely on previously shared information. Figure 22c depicts the results to the question with whom the information provider *would share* the information item. The results fit much more to the intuitive expectations: with increasing level of privacy the ratio of information items which do not get shared increases, while the other categories decline. In addition, the share of the FRIENDS category decreases,

4 The dataset does not contain enough values to allow adding both effects simultaneously.

ATTRIBUTE	SPEARMAN'S RHO	PEARSON'S R
tie strength (IS)	0.17 (p<0.05)	0.15 (p<0.05)
sympathy (IS)	0.22 (p<0.05)	0.19 (p<0.05)
social context similarity (IS)	0.10 (p<0.05)	0.10 (p<0.05)
content knowledge similarity (IS)	0.07 (p=0.09)	0.04 (p=0.27)

Table 22: Correlation of relevance and social attributes in Experiment 7a, p-values based on t-distribution

while LIMITED AUDIENCE increases – this could suggest that people control the recipients of their information items with much more rigor if the information item exceeds a certain privacy level. Figure 23 shows that the decline of the YES category (i.e., would share when explicitly asked) is noticeable with increasing degree of privacy, however, the line is much flatter than in the proactive approach shown in Figure 22c. This suggests that people are much more willing to share information when they get explicitly asked for it by the information seeker.

16.6.1.3 *RQ 2a: How Relevant Are Information Items Taken From Non-Public Information Spaces of Socially Close People when Satisfying Information Needs?*

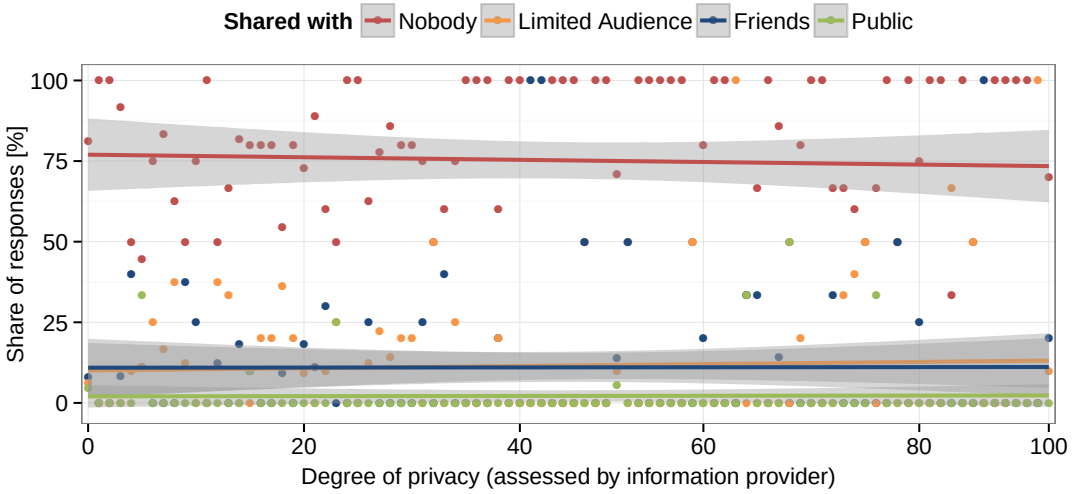
Experiment 7a provided a comprehensive dataset of 925 requests initiated from the participants. Out of those 925 requests, 648 have valid and complete values for the estimated relevance of the reply and the social attributes (tie strength, sympathy, social context similarity, content knowledge similarity) which characterize the relationship between information seeker and information provider. The reason for analyzing this dataset is to extract patterns which could help to improve the relevance for the information seeker – therefore, only data available to the information seeker is used in the analysis (i.e., all observations examined in this section are reported by the information seeker). In the following, correlations between social attributes and relevance are investigated using correlation coefficients and regression techniques.

CORRELATION COEFFICIENTS FOR RELEVANCE AND SOCIAL ATTRIBUTES Table 22 lists the correlation coefficients for the social attributes and relevance. A statistically significant positive correlation with relevance can be shown for sympathy, tie strength, and social context similarity.

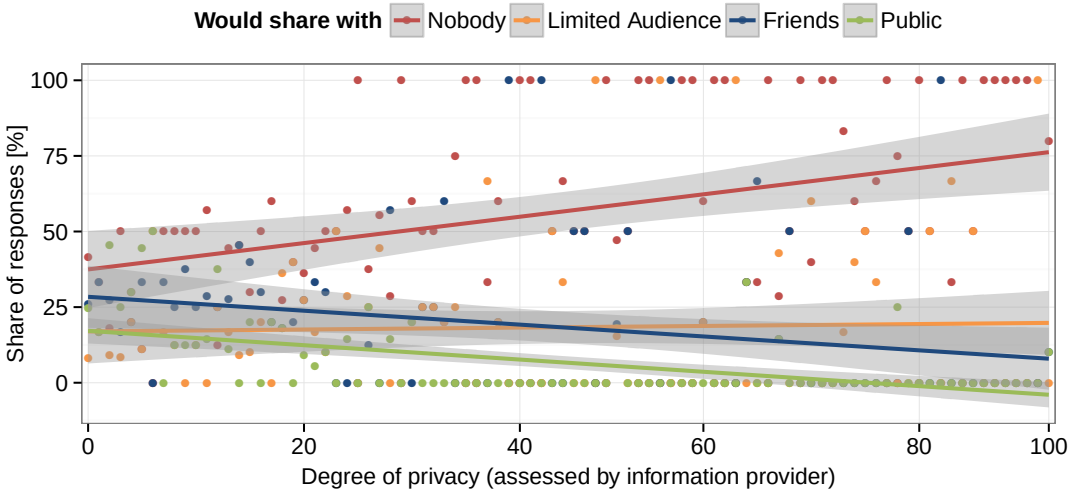
LINEAR REGRESSION MODEL TO EXPLAIN RELEVANCE WITH SOCIAL ATTRIBUTES Fitting a linear regression model to explain relevance with the social attributes tie strength, sympathy, social context similarity, and sympathy leads to the coefficients 0.1877 (sympathy, p=0.00), 0.0450 (tie strength, p=0.39), −0.0030 (content knowledge similarity, p=0.94), 0.0021 (social context similarity, p=0.96), and an intercept of 61.3406 (p=0.00). Residual standard error is 21.77 on 643 degrees of freedom, adjusted $R^2 = 0.032$ and p=0.00. Optimizing the model suggests to keep sympathy only, leading to a cleaner model with sympathy as only independent variable (coefficient 0.2210, p=0.00), an intercept of 61.9361 (p=0.00), adjusted $R^2 = 0.035$, residual standard error 21.73 on 646 degrees of freedom. Residuals of both models are not normally dis-

PRIVACY DEGREE	[0,20]	(20,40]	(40,60]	(60,80]	(80,100]
Responses	390	117	102	63	25

(a) Number of replies per privacy category (assessed by information providers)



(b) Distribution of audience categories for each level of privacy for already shared information items



(c) Distribution of audience categories for each level of privacy for information items to be hypothetically shared in the future

Figure 22: Sharing preferences for information items with different degrees of privacy (Experiment 7a)

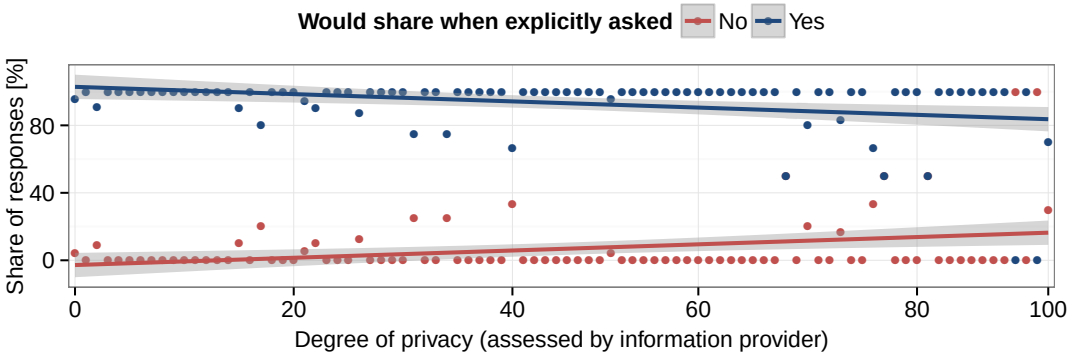


Figure 23: Information providers' willingness to share an information item when being explicitly asked for it (Experiment 7a)

VARIABLE	COEFFICIENT	STD. ERROR	DF	T VALUE	P(>|t|)
(Intercept)	66.7400	4.5140	509.2000	14.7860	0.0000
sympathy (IS)	0.1459	0.0517	589.5000	2.8190	0.0050
social context sim. (IS)	0.0098	0.0397	628.5000	0.2480	0.8041
content knowledge sim. (IS)	0.0207	0.0385	595.4000	0.5390	0.5903

Table 23: Linear regression model with random effects (information seeker, query) to explain relevance, Residual deviance: 5780.4, Null deviance: 5791.7, χ^2: 11.27, df: 3, P(>χ^2): 0.01

tributed (Shapiro-Wilk normality test), heteroscedasticity can not be shown for the full model (studentized Breusch-Pagan test, BP = 6.6377, df = 4, p-value = 0.16), but for the optimized model (BP = 7.0339, df = 1, p-value = 0.00).

LINEAR REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN RELEVANCE WITH SOCIAL ATTRIBUTES A linear regression model with random effects to account for multiple measurements of information seekers and queries reveals that the high correlation of sympathy and tie strength might negatively impact the model's quality (estimated coefficient for tie strength is 0.0308 with p-value 0.57, correlation of fixed effects with sympathy is −0.47). The parameters of an adjusted model without tie strength (only sympathy, social context similarity, and content knowledge similarity as explanatory variables) are presented in Table 23. The model suggests sympathy as the only statistically significant predictor for relevance (coef.: 0.1459, p-value: 0.00). In a similar model for tie strength (without sympathy), tie strength is not statistically significant (coef.: 0.0867, p-value: 0.07).

LOGISTIC REGRESSION MODEL TO EXPLAIN RELEVANCE WITH SOCIAL ATTRIBUTES A threshold of 100 for relevance is chosen to define the binary response variable because of the distribution of the data where a large proportion of the replies are rated as highly relevant (variable relevance, 1st qu.: 70, median: 84.5, mean: 82.2, 3rd qu.: 100; out of 648 valid replies, 203 are rated with relevance=100). The model predicts

VARIABLE	COEFFICIENT	STD. ERROR	Z VALUE	$P(>\|z\|)$
(Intercept)	-2.6049	0.4791	-5.44	0.0000
tie strength (IS)	0.0124	0.0057	2.18	0.0293
sympathy (IS)	0.0103	0.0064	1.60	0.1098
social context sim. (IS)	-0.0009	0.0042	-0.23	0.8192
content knowledge sim. (IS)	0.0009	0.0037	0.23	0.8167

Table 24: Logistic regression model to explain relevance, Residual deviance: 784.09 on 643 degrees of freedom, Null deviance: 805.71460 on 647 degrees of freedom, P(>Chi): 0.00

VARIABLE	COEFFICIENT	STD. ERROR	Z VALUE	$P(>\|z\|)$
(Intercept)	-3.2757	0.7233	-4.529	0.000
tie strength (IS)	0.0182	0.0080	2.265	0.0235
sympathy (IS)	0.0020	0.0086	0.233	0.8156
social context sim. (IS)	0.0029	0.0061	0.475	0.6344
content knowledge sim. (IS)	0.0055	0.0054	1.010	0.3124

Table 25: Logistic regression model with random effects for information seeker and query to explain relevance, Residual deviance: 690.6 on 641 degrees of freedom, Null deviance: 706.9 on 645 degrees of freedom, χ^2: 16.284 on 4 df, P(>χ^2): 0.00

the relevance category using the four social attributes as explanatory variables. In comparison to the null model, the model reduces residual deviance significantly (p-value: 0.00); the coefficients are listed in Table 24. A positive effect of tie strength is statistically significant (0.0124, p-value: 0.03).

LOGISTIC REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN RELEVANCE WITH SOCIAL ATTRIBUTES Selecting information seeker and query as random effects and the same threshold for relevance as above, the model reveals a significant positive effect of tie strength (Table 25). Creating isolated models for each explanatory variable,

- tie strength (coef.: 0.0223, std. Error: 0.0061, z value: 3.683, p-value: 0.00, residual deviance: 692.2),
- sympathy (coef.: 0.0171, std. Error: 0.0072, z value: 2.381, p-value: 0.02, residual deviance: 701.0),
- social context similarity (coef.: 0.0132, std. Error: 0.0052, z value: 2.552, p-value: 0.01, residual deviance: 700.1), and
- content knowledge similarity (coef.: 0.0108, std. Error: 0.0051, z value: 2.108, p-value: 0.04, residual deviance: 702.3)

show statistically significant positive effects on relevance.

ORDINAL LOGISTIC REGRESSION MODEL TO EXPLAIN RELEVANCE WITH SOCIAL ATTRIBUTES Splitting relevance in five equally sized categories ([0,20], (20,40], (40,60], (60,80], (80,100]) leads to bins with 19, 24, 40, 186, and 379 items. The ordinal logistic regression model fitted to the data and optimized with AIC uses *sympathy* as the only explanatory variable, with 0.0188 (p=0.00) as estimated coefficient and significantly outperforms the null model (p=0.00). According to the model, each increase in sympathy increases the odds for the next higher relevance category by factor $\exp(0.0188) = 1.0189$. However, when creating individual models, a positive effect on relevance is statistically significant for

- tie strength (coef: 0.0106, std. Error: 0.0034, z value: 3.133, p-value: 0.00) and
- social context similarity (coef: 0.0074, std. Error: 0.0030, z value: 2.483, p-value: 0.01)

while content knowledge similarity is not statistically significant (p-value: 0.47).

ORDINAL LOGISTIC REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN RELEVANCE WITH SOCIAL ATTRIBUTES A model which estimates the relevance category using the four social attributes, enhanced with random effects for "information seeker" and "query" (intercept) confirms the previously identified effect for sympathy (coef.: 0.0155, std. error: 0.0062, z value: 2.525, p-value: 0.01). The other social attributes are not statistically significant; the model is better than the null model (p-value: 0.03).

SUMMARY Table 26 shows a summary for Research Question 2a in Experiment 7a. Homophily (cf. Section 2.4.2, (Tang et al., 2014)) would suggest that people with close social ties have a higher chance for a larger overlap of common interests and therefore could possibly provide relevant information to each other, leading to higher relevance judgments for information from socially close information providers. While logistic regression and correlation coefficients confirm such an effect, it is not possible to observe it in the other models. The logistic regression model with random effects suggests that each increase in tie strength would increase the odds for the higher relevance category (with relevance = 100) by a multiplicative factor of 1.01 (without random effects) and 1.02 (with random effects). The effect is rather small and given the fact that it was not possible to observe it in the ordinal regression models, it could also be caused by the way relevance values were split in low and high categories. To summarize, our findings indicate that there might be an effect as expected and derived from literature, however, the collected data does not provide sufficient evidence to definitely prove its existence in general.

16.6.1.4 *RQ 2b: Does Social Context Imply a Valuable Contribution to Retrieving Information From the Unconscious Information Need (Serendipitous Information)?*

For the following analyses, serendipity is defined as the product of usefulness of the result ("satisfaction of information need") and the degree of unexpectedness – it represents the idea of the "lucky accident" (multiple definitions have been used in the literature, cf. Section 2.3.5). In the following, the relation between serendipity and the

VARIABLE	TIE STRENGTH	SYMPATHY	SOCIAL CONTEXT	CONTENT KNOWLEDGE
Correlation coefficients	↑	↑	↑	×
Linear regression	×	↑	×	×
Linear regr. w/ random effects	×	↑	×	×
Logistic regression	↑	×	×	×
Logistic regr. w/ random effects	↑	×	×	×
Ordinal logistic regression	×	↑	×	×
Ordinal log. regr. w/ random effects	×	↑	×	×

Table 26: Summary: Impact of social attributes tie strength, sympathy, social context similarity, and content knowledge similarity on relevance; ↑ (↓) indicates a significant positive (negative) impact, × indicates no correlation (Experiment 7a)

ATTRIBUTE	SPEARMAN'S RHO	PEARSON'S R
tie strength (IS)	-0.03 (p=0.48)	-0.00 (p=0.96)
sympathy (IS)	-0.03 (p=0.41)	-0.02 (p=0.62)
social context similarity (IS)	0.02 (p=0.66)	0.06 (p=0.15)
content knowledge similarity (IS)	0.06 (p=0.11)	0.10 (p=0.01)

Table 27: Correlation of serendipity and social attributes in Experiment 7a, p-values based on t-distribution

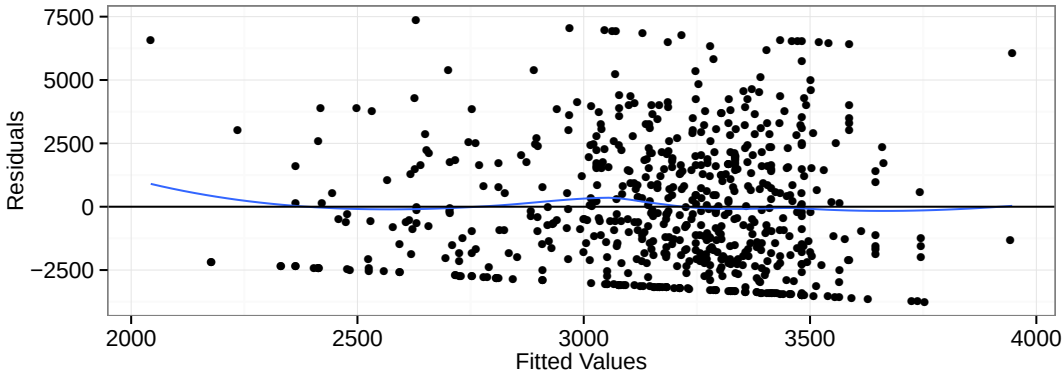
gathered social attributes is investigated using correlation coefficients and regression models (linear, logistic, and ordinal logistic).

CORRELATION COEFFICIENTS FOR SERENDIPITY AND SOCIAL ATTRIBUTES Table 27 outlines the correlation coefficients for serendipity and the social attributes tie strength, sympathy, social context similarity, and content knowledge similarity. The data does not reveal statistically significant results using Spearman's rho (all p-values $\geqslant 0.05$); for Pearson's r, content knowledge similarity is suggested to be positively correlated with serendipity (0.10, p-value: 0.01).

LINEAR REGRESSION MODEL TO EXPLAIN SERENDIPITY WITH SOCIAL ATTRIBUTES The parameters of a linear regression model to explain serendipity as a function of tie strength, sympathy, social context similarity, and content knowledge similarity are shown in Figure 24a. Figure 24b shows the residuals against the fitted values – as in previous models, the residuals are skewed because of the boundaries of the data (unexpectedness and satisfaction are in the interval $[0, 100]$, causing serendipity to be in the interval $[0, 10000]$). An optimization step based on AIC estimates 9.664 as coefficient for content knowledge similarity (std. error: 3.947, p=0.01) and an intercept of 2530.608 (p=0.00). Residual standard error is 2546 on 646 degrees of freedom, multiple R-squared is 0.0092 and adjusted R-squared is 0.0077. The F-statistic is 5.995 on 1 and 646 degrees of freedom (p-value of 0.015). The residuals are not distributed

VARIABLE	COEFFICIENT	STD. ERROR	T VALUE	P(>|t|)
(Intercept)	2908.7855	482.2622	6.03	0.0000
tie strength (IS)	-5.2088	6.1181	-0.85	0.3949
sympathy (IS)	-4.6437	6.6511	-0.70	0.4853
social context sim. (IS)	6.0219	4.8744	1.24	0.2171
content knowledge sim. (IS)	9.5659	4.3429	2.20	0.0280

(a) Residual standard error: 2546 on 643 degrees of freedom; Adjusted R-squared: 0.0072; F-statistic: 2.166 on 4 and 643 degrees of freedom; p-value: 0.07



(b) Residuals against fitted values

Figure 24: Linear model to explain serendipity with social attributes (Experiment 7a)

normally in both models, so the model's explanatory power should be considered with care (Shapiro-Wilk normality test).

LINEAR REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN SERENDIPITY WITH SOCIAL ATTRIBUTES The coefficients and statistics of a random effects model considering information seeker and query as random effects are shown in Table 28. No explanatory variable is statistically significant at $p = 0.05$ (content knowledge similarity is close). The model improves the null model (Null deviance: 11,968, Residual deviance: 11,961, χ^2: 6.8439, p-value: 0.14), but the effect is statistically not significant. An additional model with content knowledge similarity as the only explanatory variable (coef.: 8.263, p-value: 0.05) does not confirm the positive impact of content knowledge similarity on serendipity at $\alpha = 0.05$ (although it is close).

LOGISTIC REGRESSION MODEL TO EXPLAIN SERENDIPITY WITH SOCIAL ATTRIBUTES Transforming the serendipity values into two groups, one with high serendipity (> 5000, 152 observations) and one with low serendipity ($\leqslant 5000$, 496 observations) allows to use logistic regression to build a model of the data. When fitting and optimizing the model, the only explanatory variable left in the model is content knowledge similarity with a coefficient of 0.0070 (std. Error: 0.0038) at a statistically not significant level (p=0.07).

VARIABLE	COEFFICIENT	STD. ERROR	T VALUE	P(> t)
(Intercept)	2992.052	531.730	5.627	0.0000
tie strength (IS)	-8.165	6.326	-1.291	0.1972
sympathy (IS)	-2.391	6.793	-0.352	0.7250
social context sim. (IS)	6.514	5.061	1.287	0.1985
content knowledge sim. (IS)	8.904	4.598	1.937	0.0532

Table 28: Linear model with random effects (information seeker, query) to explain serendipity with social attributes, Null deviance: 11,968, Residual deviance: 11,961, χ^2: 6.8439, Degrees of freedom: 4, p-value: 0.14 (Experiment 7a)

LOGISTIC REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN SERENDIPITY WITH SOCIAL ATTRIBUTES Considering information seeker and query as random effects, a logistic regression model to explain serendipity with the social attributes and the same threshold as in the normal logistic regression model does not contain any statistically significant coefficients. Building an individual model with content knowledge similarity as single explanatory variable (given its relatively low p-value in the full model) does not lead to any additional statistically significant insights.

ORDINAL LOGISTIC REGRESSION MODEL TO EXPLAIN SERENDIPITY WITH SOCIAL ATTRIBUTES For the model, serendipity was transformed to a 3-level ordinal scale ([0, 3330], (3300, 6670], (6670, 10000]), matching a low / medium / high logic (splitting serendipity in five categories does not lead to statistically significant coefficients for models with and without random effects; random effects: information seeker, query). The observations are distributed as follows: 370 ([0,3330]), 210 ((3300,6670]), 68 ((6670,10000]). The fitted and optimized model (first degree polynomials) only has content knowledge similarity as explanatory variable (coefficient: 0.0072, std. Error: 0.0031, p=0.02). This can be interpreted in the following way: each increase in content knowledge similarity influences the odds for the higher serendipity category by a multiplicative factor of $\exp(0.0072) = 1.01$. The model outperforms the null model (p=0.02).

ORDINAL LOGISTIC REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN SERENDIPITY WITH SOCIAL ATTRIBUTES A model with all four social attributes and random effects for information seeker and query confirms the positive effect of content knowledge similarity (coef: 0.0085, std. error: 0.0042, z value: 2.052, P(> |z|)=0.04); the other explanatory variables are not significant). However, the model does not outperform the null model (p=0.27).

SUMMARY Table 29 shows a summary for Research Question 2b in Experiment 7a. According to the fitted models, the social attributes tie strength, sympathy, and social context are not related to serendipity on a statistically significant level. Only content knowledge similarity appears to have a slightly positive effect on serendipity (at least in Pearson's correlation coefficient, linear regression, ordinal logistic regression).

VARIABLE	TIE STRENGTH	SYMPATHY	SOCIAL CONTEXT	CONTENT KNOWLEDGE
Correlation coefficients	×	×	×	↑
Linear regression	×	×	×	↑
Linear regr. w/ random effects	×	×	×	×
Logistic regression	×	×	×	×
Logistic regr. w/ random effects	×	×	×	×
Ordinal log. regression	×	×	×	↑
Ordinal log. regr. w/ random effects	×	×	×	↑

Table 29: Summary: Impact of social attributes tie strength, sympathy, social context similarity, and content knowledge similarity on serendipity; ↑ (↓) indicates a significant positive (negative) impact, × indicates no correlation (Experiment 7a)

Again, the effect is rather small, but it supports the idea that serendipity is based on prior knowledge (Workman et al., 2014; Thudt et al., 2012): if information seeker and provider share a higher overlap of knowledge, finding "common ground" to relate to new knowledge is easier for both parties.

16.6.1.5 *RQ 3: Which Social Concepts Influence the Users' Routing Decisions?*

During the preparation phase of Experiment 7 (cf. Section 16.5.1), the participants uploaded 10,999 directed edges to other participants and Facebook friends with a tie strength > 0. Out of those, 689 relations have been used to send out at least one query in Experiment 7a. In the following, the differences between those two groups are analyzed in detail: what characterizes the relationships to users who got selected as information providers?

DISTRIBUTION OF SOCIAL ATTRIBUTES The median values of the social attributes tie strength, sympathy, social context similarity, and content knowledge similarity (only considering the edges where the information seeker was able to give an estimate on content knowledge similarity, cf. Section A.2.2) differ between the two groups: those edges that have been used to contact an information provider have higher median values in each social attribute (cf. Figure 25). A Wilcoxon rank sum test confirms that the two groups (selected and not selected edges) have values drawn from different populations for tie strength, sympathy, social context similarity, and content knowledge similarity (all p=0.00).

LOGISTIC REGRESSION MODEL TO EXPLAIN SELECTED RELATIONS WITH SOCIAL ATTRIBUTES A logistic regression model to explain whether a relationship is selected based on the social attributes tie strength, sympathy, social context similarity, and content knowledge similarity is detailed in Table 30. All social attributes have a statistically significant positive estimate, with tie strength having the biggest effect: one single unit increase in tie strength is predicted to increase the odds for that specific relationship to get selected by a multiplicative factor of $\exp(0.0263) = 1.0266$

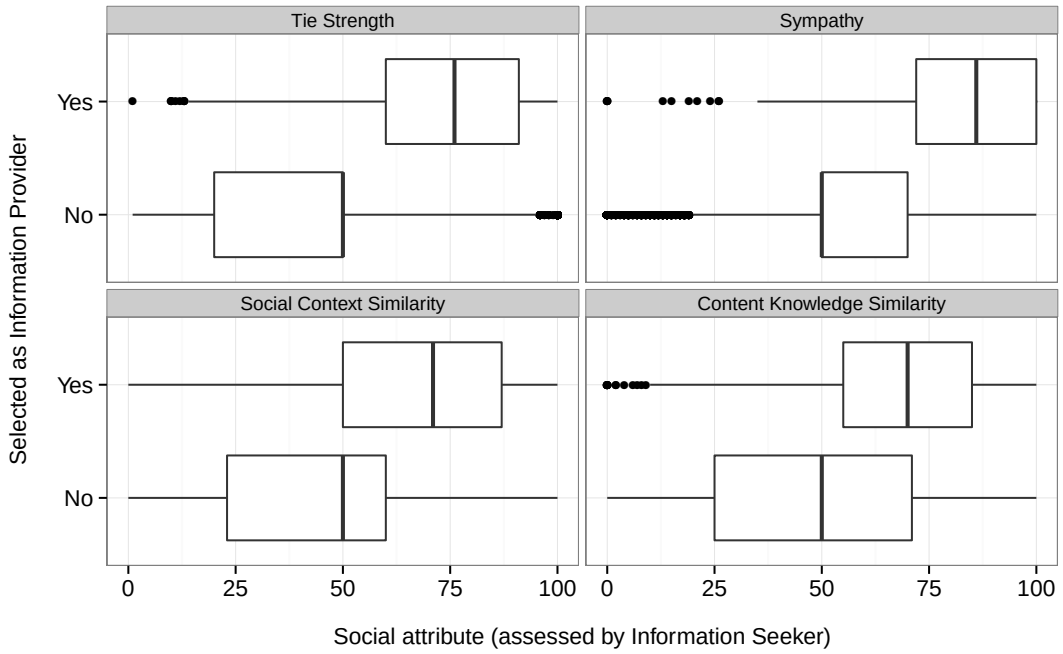


Figure 25: Social edge attributes for selected and not selected social contacts (Experiment 7a)

(when keeping the other factors constant). The model significantly improves the null model (decrease of residual deviance of 912.81 on 4 degrees of freedom, p=0.00).

LOGISTIC REGRESSION MODEL WITH RANDOM EFFECTS When considering the information seeker as random effect in the model, all social attributes positively influence the odds of an edge to get selected by the information seeker. Table 31 shows the coefficients and statistical parameters of the fitted model. The biggest effect is caused by tie strength (coef.: 0.0381, i.e. one unit increase in tie strength multiplies the odds to get selected with factor $\exp(0.0381) = 1.0388$. The model is significantly better than the null model (null deviance: 4297.8, residual deviance: 3307.4 on 4 degrees of freedom, χ^2: 990.34, p-value: =0.00).

VARIABLE	COEFFICIENT	STD. ERROR	Z VALUE	P(>|Z|)
(Intercept)	-6.4111	0.2048	-31.31	0.0000
tie strength (IS)	0.0263	0.0024	10.90	0.0000
sympathy (IS)	0.0199	0.0028	7.05	0.0000
social context similarity (IS)	0.0066	0.0019	3.41	0.0007
content knowledge similarity (IS)	0.0129	0.0017	7.45	0.0000

Table 30: Logistic regression model to explain why a social contact was chosen as an information provider in Experiment 7a; Null deviance: 4443.6 on 7245 degrees of freedom; Residual deviance: 3530.8 on 7241 degrees of freedom

VARIABLE	COEFFICIENT	STD. ERROR	Z VALUE	P(>|Z|)
(Intercept)	-7.1623	0.2703	-26.495	0.0000
tie strength (IS)	0.0381	0.0029	13.293	0.0000
sympathy (IS)	0.0151	0.0033	4.538	0.0000
social context similarity (IS)	0.0073	0.0022	3.309	0.0009
content knowledge similarity (IS)	0.0168	0.0020	8.558	0.0000

Table 31: Logistic regression model with random effects (information seeker) to explain when a social contact was chosen as an information provider in Experiment 7a; Null deviance: 4297.8, Residual deviance: 3307.4 on 4 degrees of freedom, χ^2: 990.34, p-value: =0.00

TYPES OF SOCIAL INTERACTION In addition to the edge attributes, information seekers and providers also assessed their relationship with regard to Fiske's elementary forms of social interaction, Communal Sharing (CS), Authority Ranking (AR), Equality Matching (EM), and Market Pricing (MP) (cf. (Fiske, 1992) and Section 2.4.2). For a detailed list of the questions, please refer to Table 38. Equality Matching (EM) and Market Pricing (MP) have been covered by one single question because both describe a concept of quantified, mutual benefits.

The results shown in Figure 26 indicate that the density curves for information seekers and information providers do not differ much: the AR curves are skewed to the left, indicating that there is no huge difference in authority within the group of participants which can be expected because all participants have been students without any formal hierarchy. The findings suggest that both groups are very supportive (CS curve is skewed to the right), i.e., the majority of information seekers seem to activate relations that are characterized by a high communal sharing component. Combining this finding with the answers to the EM/MP related questions also shown in Figure 26 suggests that the information providers' motives for sharing information are not based on a market idea but on (more or less) altruistic ideals. Information seekers (and to a larger extent, also information providers) do not think that providing information does induce a "social liability". In a real social information retrieval system, we would expect a different result: non-hierarchical requests may be the typical case for students but would most likely not apply to a more professional context. The large amount of CS-heavy relations might change outside a dedicated social information retrieval experiment. The same applies to the EM/MP category: the high ratio of low values for the information providers (blue curve) indicate that the information providers hesitated to explicitly state that they have the impression that information seekers owe them something (e.g., due to societal reasons). The information seekers' EM/MP curve shows a less steep curve towards 0 and suggests that a larger amount of information seekers assume that they have a liability. Alternatively, it could also be the case that the data correctly reflects the participants' opinion, suggesting that the participants follow an altruistic ideal of supporting each other without expecting anything in return.

Figure 27 shows the difference of the information seeker's rating and the rating given from the information provider for each query/response pair. A value around

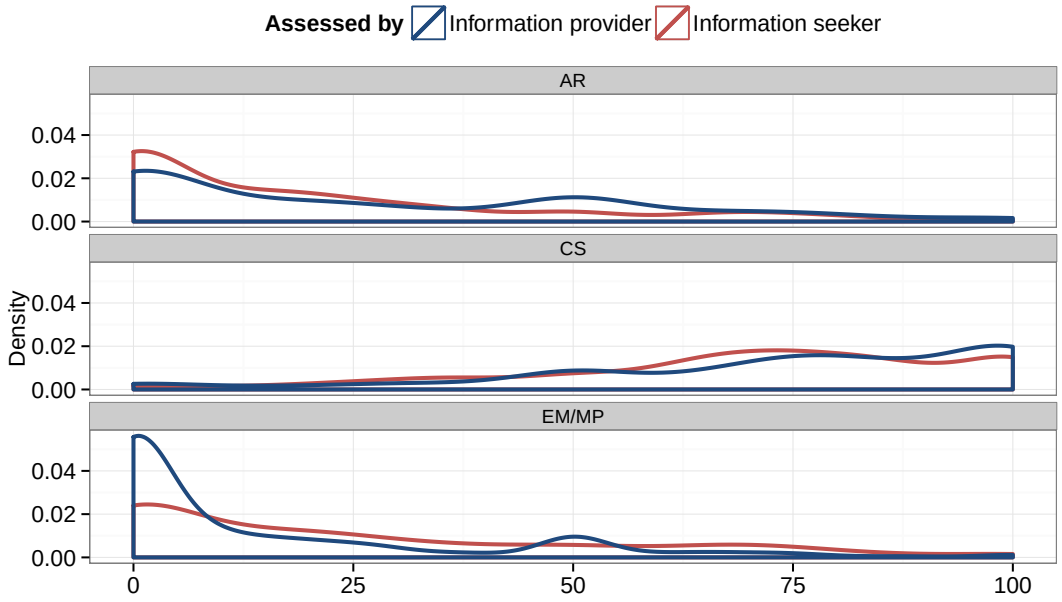


Figure 26: Density of Fiske's elementary forms of sociality: Authority Ranking (AR), Communal Sharing (CS), and Equality Matching (MP)/Market Pricing (MP) (Experiment 7a)

0 indicates that both gave roughly the same estimate, a negative value documents that the information provider gave a higher estimate on average and vice versa. In the majority of the cases, the information seeker/providers estimated the dimensions similarly, since all three curves are distributed around 0. In the AR component, the information providers reported by tendency slightly higher values, while the CS curve is even more balanced. In the plot covering the EM/MP components it is notable that information providers gave lower estimates than information seekers. This could be caused by the fact that the information providers did not feel comfortable to assert a claim against the information seekers (even if it was made clear that the response would not be revealed to anyone, especially not the other party).

RELATION TO SOCIAL CAPITAL MODEL FOR INFORMATION MARKETS In Section 7.1.3, a market model for social capital was introduced. Even if a full evaluation has not been conducted as part of this thesis, the ascertained data of Experiment 7a can be discussed against the background of the proposed market model. The initial participants of Experiment 7a (i.e., the students) were asked to nominate potential information providers for three individually defined and three predefined queries. In terms of the proposed model, for each question, each student A had to find a social contact V for whom the following equation is positive:

$$(G^{IS}_{A,V} - C^{IS}_{A,V}) + C^{IP}_{A,V} - G^{IP}_{A,V}|V + AD_{A,V} \geqslant 0 \tag{69}$$

$$G^{IS}_{A,V} - G^{IP}_{A,V}|V + AD_{A,V} \geqslant C^{IS}_{A,V} - C^{IP}_{A,V} \tag{70}$$

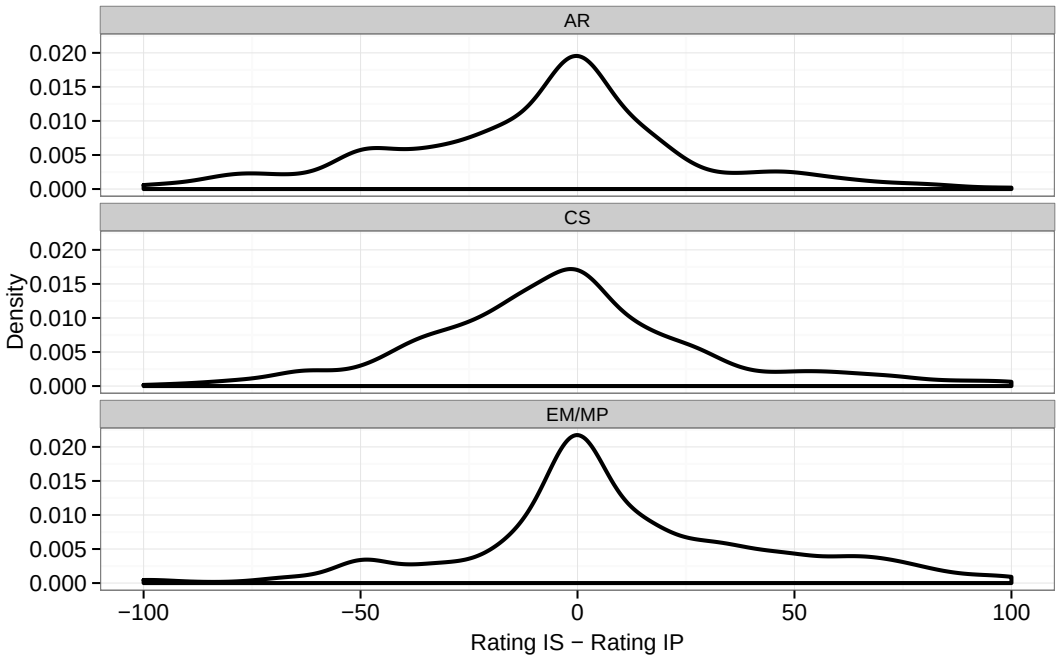


Figure 27: Rating differences for Fiske's elementary forms of sociality (rating of information seeker - information provider for each individual query/response pair in Experiment 7a)

A would try to find someone who helped A in the past with high-quality answers (high value of $G^{IS}_{A,V}$) or who received a lot of help from A in the past ($C^{IP}_{A,V}$). At the same time, the costs for interacting with this person should not be too high ($C^{IS}_{A,V}$ should be minimized, especially when considering that the whole scenario takes place as an experiment that is part of a lecture). When analyzing which contacts got selected, it is obvious that the participants tended to select people with high tie strength and sympathy (cf. Figure 25). This could be explained by the fact that the anticipated costs to contact them are lower while the expected goodwill in their decision equations as information provider and the resulting audacity parameter are higher. Figure 26 and Figure 27 also highlight two things:

- The participants activated particularly those social relations with large CS components (or explicitly asking caused a bias in the data) (Figure 26).
- Both parties (information seeker and provider) agreed on the composition of their respective social relation (Figure 27). While some information providers felt obliged to answer the question (slight plateau on the negative side of the scale in Figure 26 for AR), more information seekers tend to feel that they owe something to the information provider (EM). The CS dimension was rather balanced.

These results also suggest that information seekers tend to select information providers who would assess the interaction in a similar way.

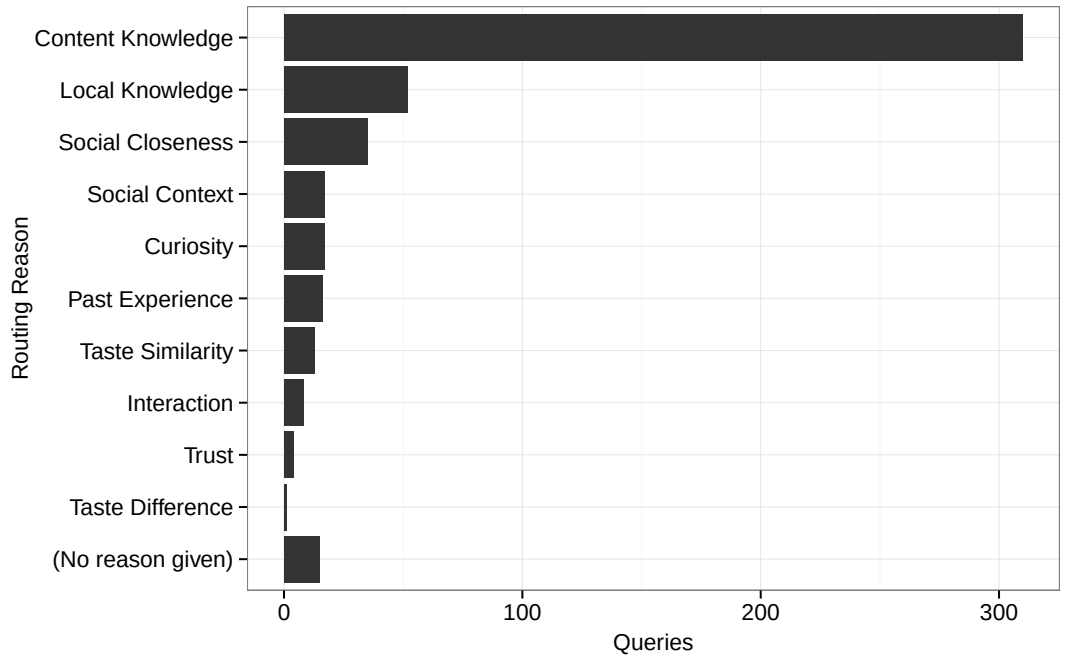


Figure 28: Frequency of reasons for routing a query to a specific information provider (Experiment 7a)

REASONS FOR CHOOSING A SPECIFIC INFORMATION PROVIDER When submitting a query, the information seekers have been asked to specify why this person would be a good information provider (cf. Table 38). An analysis of responses to the self-defined queries revealed the clusters listed in Table 32, the frequency of each category is shown in Figure 28. The results need to be interpreted carefully and can not be generalized as-is, because the underlying set of queries is not controlled (each information seeker could specify queries independently, the predefined queries have not been considered to avoid a bias based on the type of predefined queries). In a more controlled experiment, the same set of predefined queries covering a broad range of content areas would have been routed by each information seeker to an information provider, to avoid any prior probabilities for queries and/or interaction preferences. Nevertheless, this analysis is useful to give an indication about the most recurring reasons for routing decisions, since it includes the prevalence rates of query types and content categories. The (by far) most common reason to select a person as potential information provider is her CONTENT KNOWLEDGE (310), followed by LOCAL KNOWLEDGE (52) and SOCIAL CLOSENESS (35). TASTE DIFFERENCE (1), TRUST (4), and INTERACTION (8) have been rarely mentioned reasons. Thus, a routing algorithm focussing on content knowledge (as e.g. detailed in Section 7.1.2) could successfully simulate the routing decision of a human information seeker in a majority of the cases.

category	description	example query / routing reason
Content Knowledge	The IP is considered as knowledgeable in the query's content domain.	*"Could a computer have a mind?" / "He is doing research about Neuroscience and he has posted many topics about these things on his social networks"*
Local Knowledge	The IP is considered as a local expert.	*"How much are monthly living expenses on average in Canada?" / "This recipient lives and works in Canada, therefore I think I can receive a satisfying answer for my question."*
Social Closeness	IP and IS are socially close.	*"What advice can you give me to be more relaxed when meeting other people?" / "He is a close friend and even though he is not the most relaxed person in the world he is mature."*
Taste Similarity	IP and IS share the same taste.	*"What are must-read-books?" / "We often share the same taste according to books."*
Social Context	IS and IP share the same social context.	*"What to buy my girlfriend for birthday present?" / "He knows my girlfriend."*
Past Experience	IS wants to benefit from IP's past experience.	*"Do you think that Rock Am Ring is a good festival?" / "He was at Rock im Park 2015."*
Curiosity	IS is curious what the IP thinks about the respective topic.	*"If you are given a chance to be in the body of someone else, who would you pick?" / "He is very innovative to form a good answer for such question."*
Interaction	IP has been chosen primarily to start/maintain interaction between IS and IP.	*"How much do you weigh?" / "We duel ourselves about our weight on a daily basis."*
Trust	IS trusts in IP's opinion.	*"Should I keep the beard that way?" / "I value her opinion."*
Taste Difference	IS chooses IP because of her different taste	*"Which movie would you recommend to watch with a close friend?" / "She doesn't watch the same movies that I do so I could discover new ones."*
N/A	No reason given	– / –

Table 32: Information seekers' reasons for routing a query to its respective information provider in Experiment 7a

16.6.1.6 *RQ 4: Which Categories of Information Needs Could Benefit From Social Information Retrieval?*

To identify which types of information needs and content areas are suitable for social search, the subset of user-defined queries was analyzed (324 unique queries). These queries have been mapped to content and type categories. In a 3-step process, each query was assigned to a type category (based on Oeldorf-Hirsch et al.'s categories (Oeldorf-Hirsch et al., 2014)) and a content category (from a self-defined content category system) by two independent raters. After the first rating round, Cohen's kappa (a measure for inter-rater agreement, (Cohen, 1960)) was 0.43 (type) and 0.72 (content). With a follow-up discussion and a subsequent, more precise definition of the categories, Cohen's kappa could be increased to 0.99 (type) and 0.97 (content). This result can be interpreted as substantial agreement between the raters. The classification is part of Carola Boettcher's Bachelor's Thesis (Boettcher, 2016), which was supervised by Christoph Fuchs and Georg Groh at the Chair for Applied Informatics – Cooperative Systems at Technische Universität München. The three predefined queries (Section 16.5.2) have not been considered in this analysis to avoid a bias towards recommendation. It is also important to note that the participants could see the predefined queries only after they had defined their own individual queries (i.e., the predefined queries did not influence the participants' choice).

In addition, the queries that had received an answer (and their respective content/type category) were examined further with regard to their performance for satisfaction, relevance, personalization, and degree of unexpectedness (265 unique queries; if a query has received multiple answers, the ratings were averaged for the respective query).

CONTENT CATEGORIES The content categories and their definitions are listed in Table 33. Figure 29 shows the distribution of unique questions on the content categories. MEDIA (54), TECH (53), FOOD (46), TRAVEL (42), and LIFESTYLE (36) were the most commonly chosen topics, whereas SOCIAL (5), SCIENCE (5), SPORTS (7), PERSONAL (7), and SHOPPING (9) were the scarcest content categories.

Figure 30 shows the distribution of satisfaction, relevance, personalization, and unexpectedness for each content category, rated by the information seekers. High median values for satisfaction were achieved in content areas like SOCIAL (100), SCIENCE (91), FOOD (90.25), and TECH (87). Only in the SCIENCE category, the first and third quartile of the data (i.e., the box) reached into the area of 50 and below. Content areas with high median relevance values included SOCIAL (92), MEDIA (89.5), FOOD (89.5), SHOPPING (88.5), and TECH (88.5). Again, only SCIENCE had parts of the replies in quartile 1-3 in the area of relevance < 50. The results for personalization were low, apart from IDEOLOGY (70) and JOB (65.25), all categories had median values below 50. SPORTS (12.25), SCIENCE (24), and SOCIAL (26) were the areas with the lowest median degree of personalization. When looking at the degree of unexpectedness, the categories JOB (55.33) and SOCIAL (52.5) were the only ones that have median values above 50, the categories with the lowest results are SCIENCE (22.5), SPORTS (27.5), and PERSONAL (28).

The Kruskal-Wallis rank sum test does not confirm a statistically relevant difference between the content categories with regard to the performance in satisfaction

category	description
Travel	Travel destinations, but also nice places in Munich or nearby (e.g., *"How much money does a trip to Prague cost?"*)
Media	Books, movies, computer games, etc. (e.g., *"Should I watch the second Hunger Games movie if I have liked the first one and not have read the books?"*)
Food	Restaurant, recipes, beverages, etc. (e.g., *"Can you recommend an easy to cook meal which is also tasty?"*)
Tech	Hardware, software, programming languages, etc.(e.g., *"Would you recommend using Ubuntu as OS?"*)
Education	Questions related to university, learning in general (e.g., *"Which bock can you recommend to read in order to improve my English?"*)
Job	Job applications, salaries, etc. (e.g., *"Do you know any company here in Munich related to biomedical computing and image processing so I can apply as student?"*)
Health	Physical and mental health (e.g., *"Why some medications are recommended to be taken sublingually?"*)
Social	Questions related to social interactions, e.g. birthday presents, etc. (e.g., *"You are invited to a BBQ with your friend. What do you bring with you?"*)
Shopping	Question where to buy things where the focus is on the act of buying and not on the content of the product (e.g., *"Where can I buy a cheap bike?"*)
Lifestyle	Way and/or place of living, parties/going out, concerts, etc. (e.g., *"What do you think of selfies?"*)
Science	Things that can be explained in a scientific way (e.g., *"Why does CocaCola explode after mixing it with Mentos?"*)
Ideology	Everything related to individual philosophy of life, e.g. perspectives on EU banking crisis, elections, etc. (e.g., *"Do you think Germany should financially help Greece?"*)
Sports	Football clubs, leisure sports, etc. (e.g., *"Will Hertha BSC bring Mitchell Weiser to Berlin next season?"*)
Personal	Other questions related to personal attributes like taste, perception, etc. (e.g., *"Do you usually use an alarm to wake up in the morning?"*)

Table 33: Content categories of questions asked by the participants (Experiment 7a)

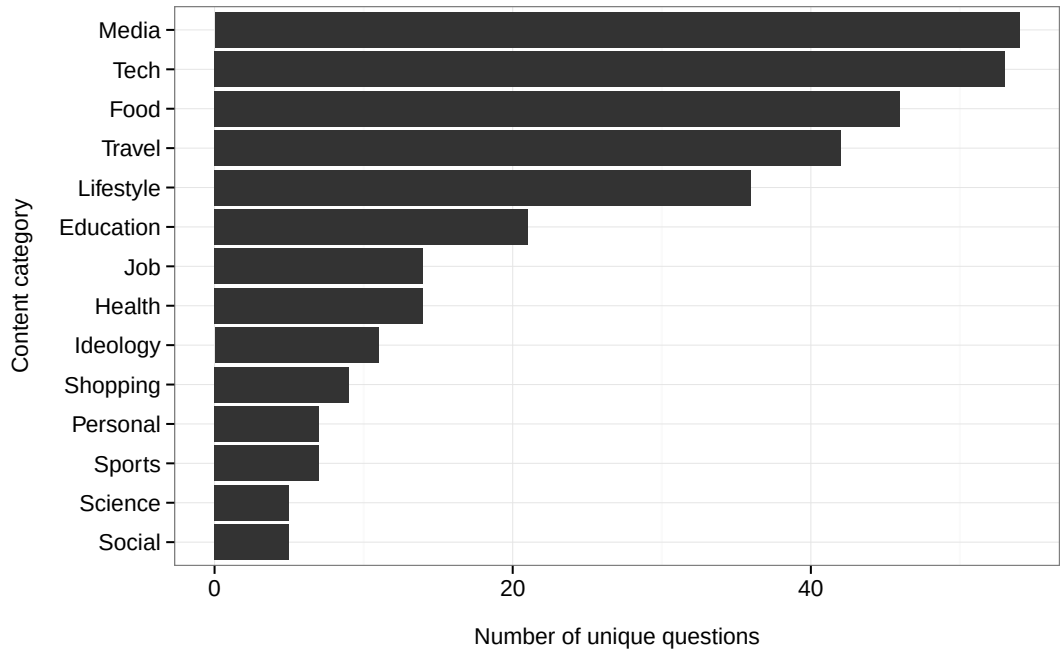


Figure 29: Number of unique queries, split by content category (Experiment 7a)

(Kruskal-Wallis χ^2 = 9.2019, df = 13, p-value = 0.76), relevance (Kruskal-Wallis χ^2 = 18.86, df = 13, p-value = 0.13), personalization (Kruskal-Wallis χ^2 = 17.045, df = 13, p-value = 0.20), and unexpectedness (Kruskal-Wallis χ^2 = 8.5841, df = 13, p-value = 0.80).

TYPE CATEGORIES To investigate the performance of social information retrieval for different types of information needs, the queries were mapped to the categories proposed by Oeldorf-Hirsch et al. (Oeldorf-Hirsch et al., 2014), extended by a category PERSONAL EXPERIENCE, which is related to FACTUAL KNOWLEDGE, but only limited to the other person's individual experience. It therefore can be seen as some sort of subjective factual knowledge.

Figure 31 shows the distribution of the queries across the type categories – the clear majority of queries is part of the RECOMMENDATION category, followed by OPINION, and FACTUAL KNOWLEDGE. Figure 32 shows how the type categories compare against each other in terms of satisfaction, relevance, personalization, and unexpectedness. The most satisfying results (based on the median value) have been achieved for queries in the categories EXPLORATORY (92), FACTUAL KNOWLEDGE (84), and RECOMMENDATION (84). While the high relevance values could have been expected to a certain degree (EXPLORATORY and RECOMMENDATION can be seen as inherently social because of their complex and subjective nature; it is also not surprising that information providers can give accurate replies to queries related to FACTUAL KNOWLEDGE – however, it is unexpected that FACTUAL KNOWLEDGE is such a common category for social information retrieval, cf. Figure 31). SOCIAL CONNECTION (0) and RHETORICAL (72.25) are the categories with the lowest satisfaction level (what might be indirectly caused by the limited number of samples in these categories, cf. Figure 31). For rel-

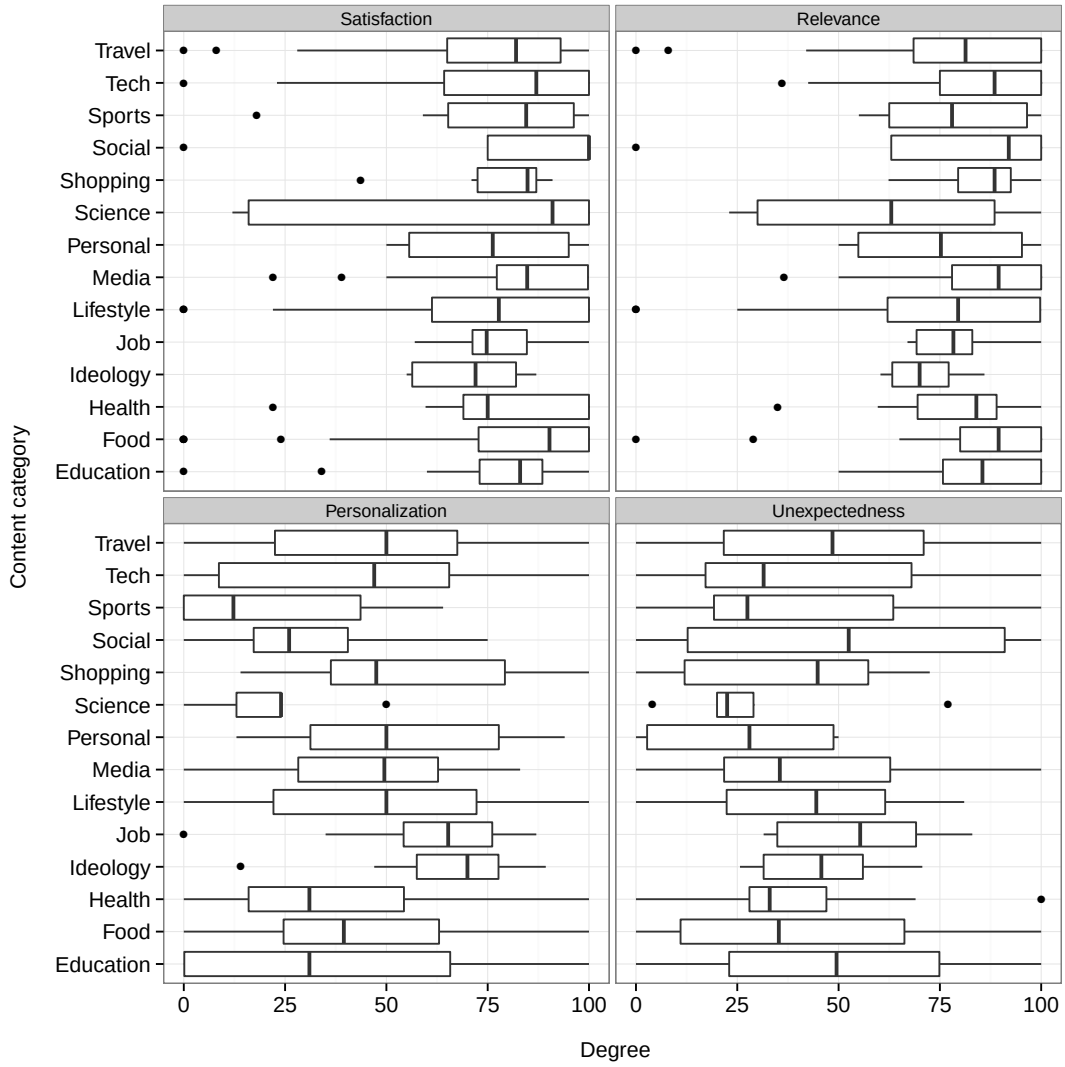


Figure 30: Satisfaction, relevance, personalization, and unexpectedness of queries, split by content category (Experiment 7a)

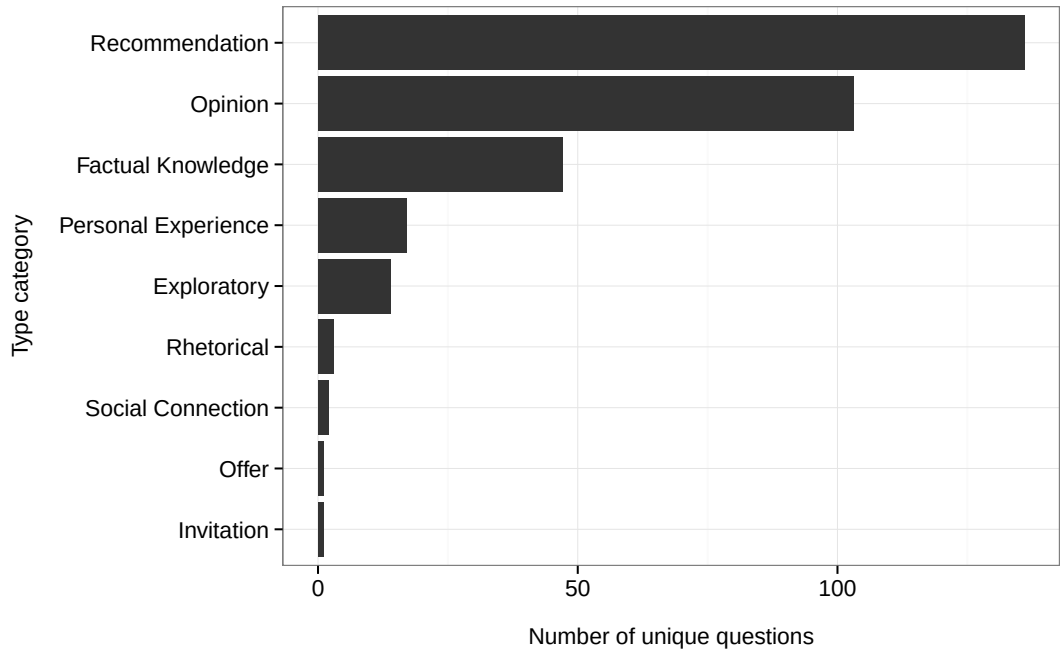


Figure 31: Number of unique queries, split by type category; categories defined by Oeldorf-Hirsch, enhanced with category PERSONAL EXPERIENCE (Oeldorf-Hirsch et al., 2014) (Experiment 7a)

evance, FACTUAL KNOWLEDGE (88), OPINION (86), and RECOMMENDATION (85) are the categories with the highest median values, whereas SOCIAL CONNECTION (0) and RHETORICAL (70.75) stay at the lowest ranks again. When looking at personalization, the categories with the highest values are PERSONAL EXPERIENCE (50), RECOMMENDATION (49), and OPINION (47). It is important to note that except PERSONAL EXPERIENCE all categories have a median degree of personalization that is below 50. The lowest values are reported for FACTUAL KNOWLEDGE (26) and SOCIAL CONNECTION (29). The queries which received the replies with the highest median value for the degree of unexpectedness are from the categories SOCIAL CONNECTION (88), PERSONAL EXPERIENCE (50), and OPINION (45.75). In contrast, queries from the categories FACTUAL KNOWLEDGE (29) and RHETORICAL (36) have received answers with a low degree of unexpectedness.

The Kruskal-Wallis rank sum test does not provide enough evidence to confirm that the performance measures differ on a statistically significant level between the type categories (satisfaction: Kruskal-Wallis $\chi^2 = 5.8548$, df = 6, p-value = 0.44, relevance: Kruskal-Wallis $\chi^2 = 5.4959$, df = 6, p-value = 0.48, personalization: Kruskal-Wallis $\chi^2 = 3.98$, df = 6, p-value = 0.68, unexpectedness: Kruskal-Wallis $\chi^2 = 5.2279$, df = 6, p-value = 0.51).

TYPE & CONTENT CATEGORIES Figure 33 shows a heat map highlighting the main categories of the questions on both axes (content, type). The most information needs fall into the area of RECOMMENDATION (FOOD, MEDIA, TECH, TRAVEL) and OPINION (TECH, LIFESTYLE, MEDIA). These domains could be interpreted as being perceived as better suited for Social Information Retrieval than other areas.

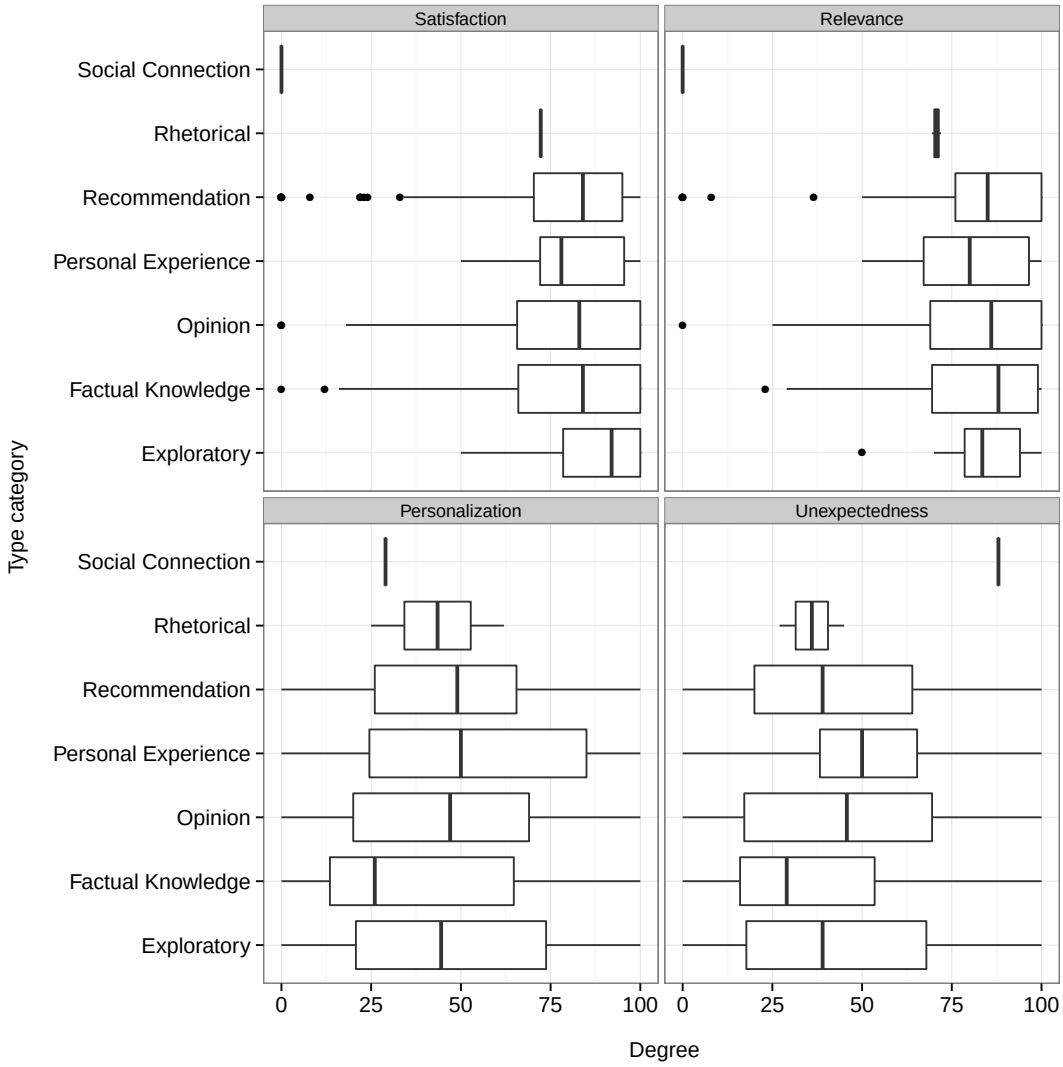


Figure 32: Satisfaction, relevance, personalization, and unexpectedness of queries, split by type category (Experiment 7a)

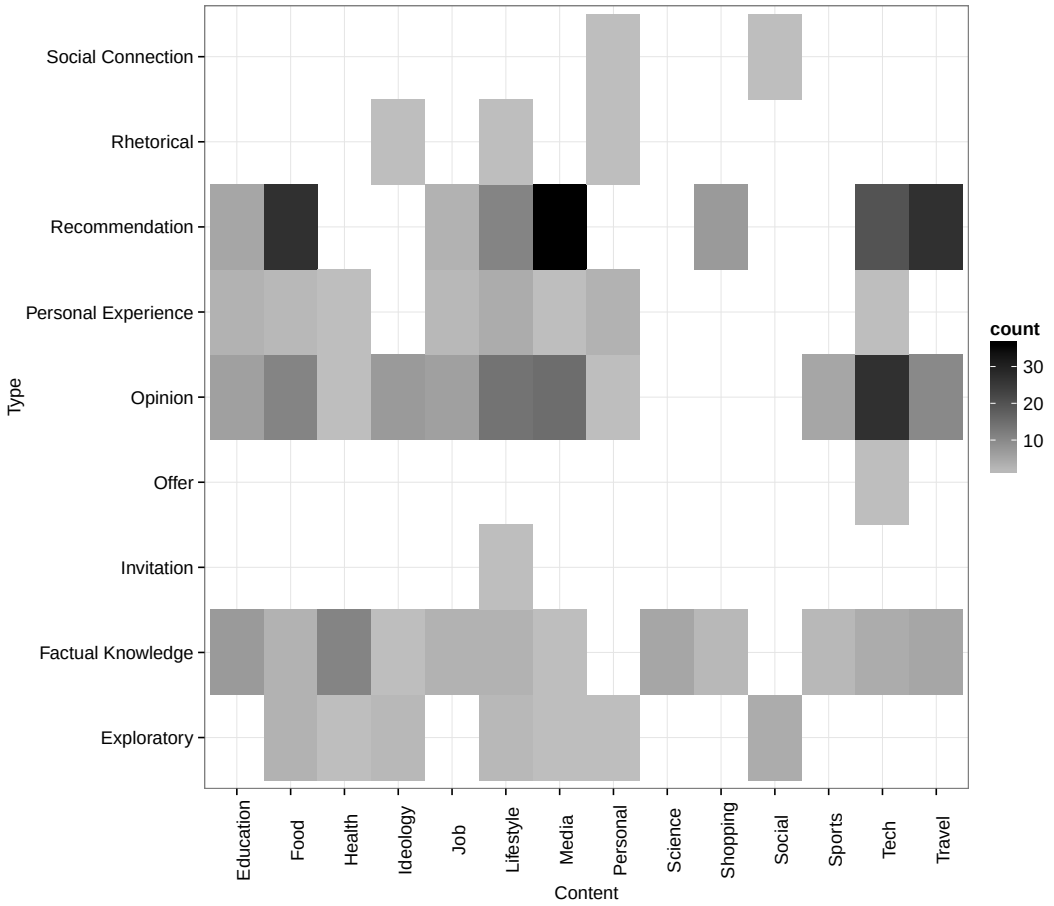


Figure 33: Distribution of unique queries, split by content and type category (manual mode)

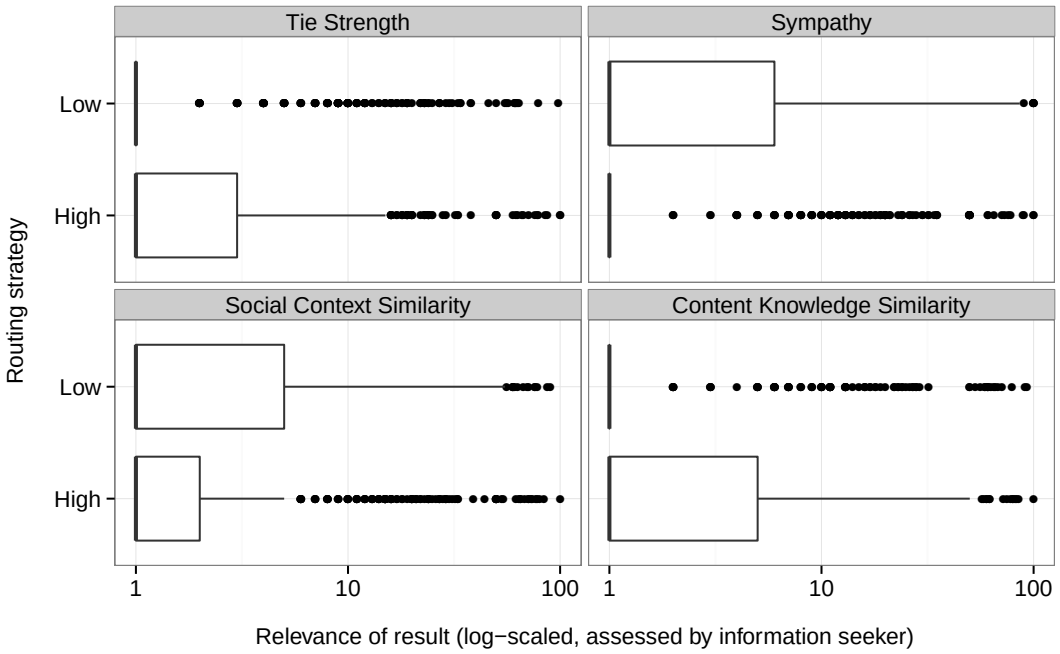


Figure 34: Relevance of replied URLs for information needs, split by strategy to select information providers (Experiment 7b)

SUMMARY Given the results of the conducted Kruskal-Wallis tests, the data does not provide a statistically sound proof to focus on one of the content or type categories in order to maximize satisfaction, relevance, personalization, or unexpectedness. Especially the content categories highly depend on the selected set of participants – e.g., their dedication to topics related to media & technology might be caused by their study background (Computer Science). Nevertheless, the two dominant type categories (RECOMMENDATION and OPINION) could constitute typical use case scenarios for social information retrieval because unlike the majority of the other type categories, the information provider's subjective information is of crucial importance for these types of information needs.

16.6.2 *Experiment 7b: Automated Social Information Retrieval Using Topic Models*

16.6.2.1 *RQ 2a: How Relevant Are Information Items Taken From Non-Public Information Spaces?*

In Experiment 7b, the recipients of a query have been selected using a randomly chosen strategy (out of eight possible strategies). The available strategies are defined using the maximum/minimum values of the social attributes, i.e. strong/weak ties, high/low sympathy, high/low social context similarity, and high/low content knowledge similarity. This data allows us to analyze performance differences between the maximum/minimum strategy for each social attribute.

DISTRIBUTION OF RELEVANCE Figure 34 shows the distribution of relevance for each strategy. One of the challenges of this dataset is that the majority of relevance

ROUTING STRATEGY	COEF. FOR RELEVANCE	P (COEF.)	P (MODEL)
tie strength (IS)	0.0190	0.03	0.03
sympathy (IS)	-0.0285	0.06	0.06
social context sim. (IS)	0.03	0.06	0.06
content knowledge sim. (IS)	0.0123	0.41	0.39

Table 34: Coefficients and significance measures of the logistic regression models to estimate low/high manifestation of the routing strategies based on social attributes using relevance (Experiment 7b)

ratings are 0. This is most likely caused by the missing overlap of the information provider's topic spaces and the information seekers' queries: topic spaces (with 100 dimensions) were not broad enough to cover the magnitude of different topics, causing a high amount of apparently irrelevant URLs to be recommended to the information seekers. In addition, the documents in the fictive information spaces (i.e., the crawled websites) differed heavily in number (and sometimes did not reflect a meaningful information space). A more detailed discussion on these peculiarities can be found in Section 16.7.

A pairwise conducted Wilcoxon Rank Sum test did not allow to reject the null hypothesis that the relevance values do not differ significantly for the strategies (tie strength: $W = 179,290$, p-value = 0.21, sympathy: $W = 122,540$, p-value = 0.12, social context: $W = 179,080$, p-value = 0.09, content knowledge: $W = 123,520$, p-value = 0.08).

LOGISTIC REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN HIGH/LOW STRATEGY TYPES WITH RELEVANCE To detect whether the low / high routing strategies of each social attribute cause different relevance results and at the same time consider differences in rating behavior for each information seeker, four logistic regression models have been fitted (one for each social attribute, i.e., tie strength, sympathy, social context similarity, and content knowledge similarity). Each model takes the relevance judgment as explanatory variable and tries to predict the respective low/high category of the social attribute. By doing this, we switch the causality of the model: instead of trying to predict the relevance caused by the different routing strategies, we focus on the question whether there is a statistically significant tendency for higher (or lower) relevance to occur in the "high" variant of the respective routing strategy. Table 34 shows the coefficient (and its p-value) and the significance of the improvement when compared with the null model for each social attribute. The results suggest that higher relevance increases the odds that the routing strategy "strong ties" has been chosen (instead of "weak ties"). This result does not confirm Granovetter's idea of the "strength of weak ties" (Granovetter, 1973; Zhang and Ackerman, 2005; Rogers, 1983) but leads to a different direction, supporting the relevance of strong ties (Panovich et al., 2012; Hansen, 1999; Uzzi, 1997). The other models do not reveal statistically significant results.

STRATEGY	SATISFACTION	UNEXPECTEDNESS	SERENDIPITY
Tie strength (strong)	4.98	66.23	319.01
Tie strength (weak)	4.26	63.00	261.98
Sympathy (high)	8.10	46.60	450.10
Sympathy (low)	7.79	61.70	408.06
Social context similarity (high)	7.83	61.42	558.05
Social context similarity (low)	5.52	56.57	345.48
Content knowledge similarity (high)	11.31	59.68	632.73
Content knowledge similarity (low)	6.66	61.97	263.35

Table 35: Average degree of satisfaction, unexpectedness, and serendipity for each information provider assignment strategy (Experiment 7b)

16.6.2.2 *RQ 2b: Does Social Context Imply a Valuable Contribution to Retrieving Information From the Unconscious Information Need (Serendipitous Information)?*

The participants also rated the degree of unexpectedness and the satisfaction with the responses they received from the information providers. This information was used to calculate a degree of serendipity, defined as the product of degree of satisfaction and unexpectedness as already defined above for each response to a query in Experiment 7a. Table 35 shows the average degree of serendipity for each strategy. The largest differences between pairs of strategies appear for social context similarity and content knowledge similarity.

WILCOXON RANK SUM TEST To test whether the respective "high" and "low" variants of each routing parameter cause different values for serendipity, Wilcoxon rank sum test is used. The test does not require a normal distribution of the residuals and therefore is suited for the collected data. The results suggest that the two strategies based on content knowledge similarity are different to a statistically significant level ($W = 5,486.5$, p-value = 0.02), whereas the strategies based on tie strength ($W = 6,954$, p-value = 0.83), sympathy ($W = 4,810$, p-value = 0.14), and social context similarity ($W = 7,833$, p-value = 0.38) are not.

LOGISTIC REGRESSION MODEL WITH RANDOM EFFECTS TO EXPLAIN HIGH/LOW STRATEGY TYPES WITH SERENDIPITY As already done before for relevance in Section 16.6.2.1, serendipity is used to estimate the odds for the "high" category for each pair of routing strategies (tie strength, sympathy, social context similarity, content knowledge similarity). Due to the fact that the assessment of *satisfaction* was only done for each response consisting of five URLs (in contrast to the assessment of relevance, which was done for each URL individually), the number of observations is not sufficient to allow the random effect models to converge (considering the information seeker as random effect). Normal logistic regression models confirm the finding from above (serendipity is considered as a positive coefficient in the model to explain whether the high or low content knowledge similarity routing strategy has

been selected). The effect is quite low (coef.: 0.0003, std. error: 0.0001), but significant (z value: 2.092, p-value: 0.04).

16.6.3 *Experiment 7c: Social Product Search*

16.6.3.1 *RQ 2a: How Relevant Are Information Items Taken From Non-Public Information Spaces?*

SIMILARITY OF PRODUCTS In a shopping scenario, one of the basic problems to solve is to decide which item to analyze further or to buy. We assume that a socially close person can help to decide when she already faced the decision herself. Thus, this research question can get reframed to investigate the correlation of buying or browsing decisions and the social attributes of the participants, i.e. whether the lists of products that were bought/viewed and the social attributes (tie strength, sympathy, social context similarity, content knowledge similarity) of two users correlate.

Each participant provided two lists with product names, one list containing all items the user bought on Amazon, the other one listing all items that the user analyzed on Amazon. Formally, each user u provided a vector $\vec{b}_u$ of bought items and a vector $\vec{v}_u$ of viewed items. The vector space dimensions are defined using Amazon's ASIN (Amazon Standard Identification Number) system (cf. Chapter 4) – if a product has been bought by u at least once, the respective dimension is set to 1 in $\vec{b}_u$ (same for $\vec{v}_u$). The similarity of vectors obtained from two users can be calculated using cosine similarity.

During the preparation phase of Experiment 7c, all participants assessed tie strength, sympathy, social context similarity, and content knowledge similarity for each other participant. Figure 35 shows the correlation of the similarity of product vectors for viewed products and social attributes, using Pearson's r and Spearman's rho, Figure 36 shows the same for the bought vectors. The social attributes reflect the information seekers' perspective.

It is notable that tie strength, sympathy, and social context similarity are correlated with the degree of similarity of the lists of viewed products, especially when focusing only on those observations which have at least one single similar product (i.e., the similarity value is > 0, see rows 2 and 4 in the table in Figure 35). This indicates that socially close people tend to be interested in the same products.

For the bought products, the Pearson coefficient does not reach a statistically significant level (p-value, cf. Section 5.1.1). Spearman's rho is close to 0 for all cases (and statistically significant for the complete dataset). When focusing on the subset where at least one common product exists between two users, the coefficient turns negative for tie strength and is not significant for the other values. Figure 36 provides the detailed data for each attribute and method and shows the plots for the social attributes and the similarity of the vectors of bought products (focusing on observations where at least one product is shared between two users).

IMPACT OF PRODUCT ORIGIN ON USEFULNESS The participants were asked to search for products they would consider buying. The query was forwarded to Amazon to obtain a result set of three items. In addition, two items were added (one that has been bought and one that has been viewed by a specific group of other users; in

TYPE	TIE STRENGTH	SYMPATHY	SOCIAL CONTEXT	CONTENT KNOWLEDGE
Pearson's r	0.078 (p=0.00)	0.068 (p=0.00)	0.054 (p=0.00)	0.063 (p=0.01)
Pearson's r [sim>0]	0.354 (p=0.00)	0.389 (p=0.00)	0.317 (p=0.00)	0.150 (p=0.27)
Spearman's rho	0.034 (p=0.00)	0.044 (p=0.00)	0.020 (p=0.04)	0.087 (p=0.00)
Spearman's rho [sim>0]	0.141 (p=0.02)	0.159 (p=0.01)	0.179 (p=0.00)	0.019 (p=0.89)

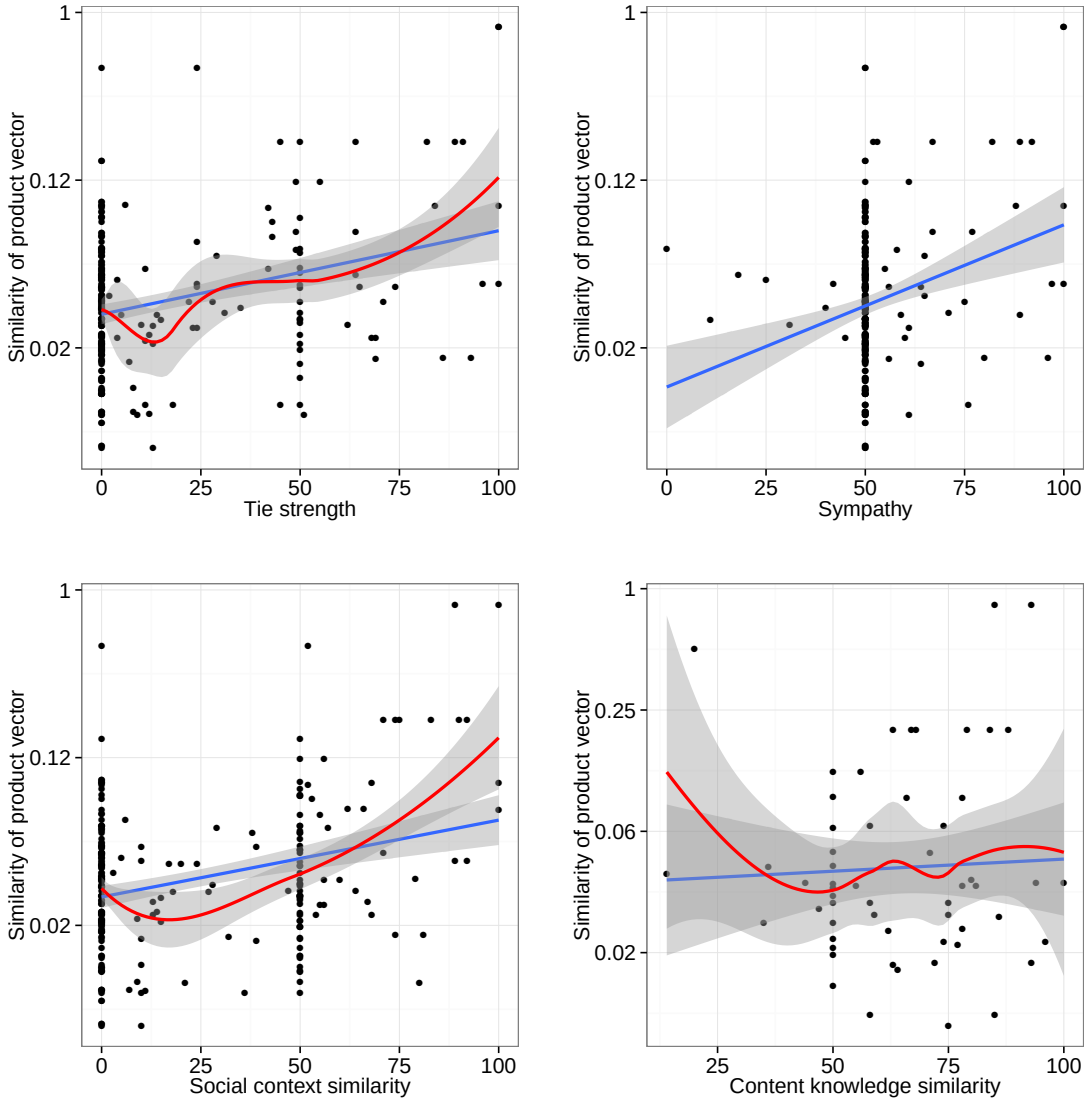


Figure 35: Correlation of social attributes and similarity of *viewed products* on Amazon; plots show data for sim>0; blue line: linear regression, red line: LOESS regression (not available for sympathy); log-scale; gray area denotes 95% confidence interval

TYPE	TIE STRENGTH	SYMPATHY	SOCIAL CONTEXT	CONTENT KNOWLEDGE
Pearson's r	-0.003 (p=0.79)	0.006 (p=0.59)	0.002 (p=0.82)	0.054 (p=0.05)
Pearson's r [sim>0]	-0.040 (p=0.35)	-0.041 (p=0.34)	-0.035 (p=0.42)	0.061 (p=0.54)
Spearman's rho	0.033 (p=0.00)	0.048 (p=0.00)	0.048 (p=0.00)	0.049 (p=0.08)
Spearman's rho [sim>0]	-0.099 (p=0.02)	-0.011 (p=0.80)	-0.056 (p=0.19)	0.170 (p=0.08)

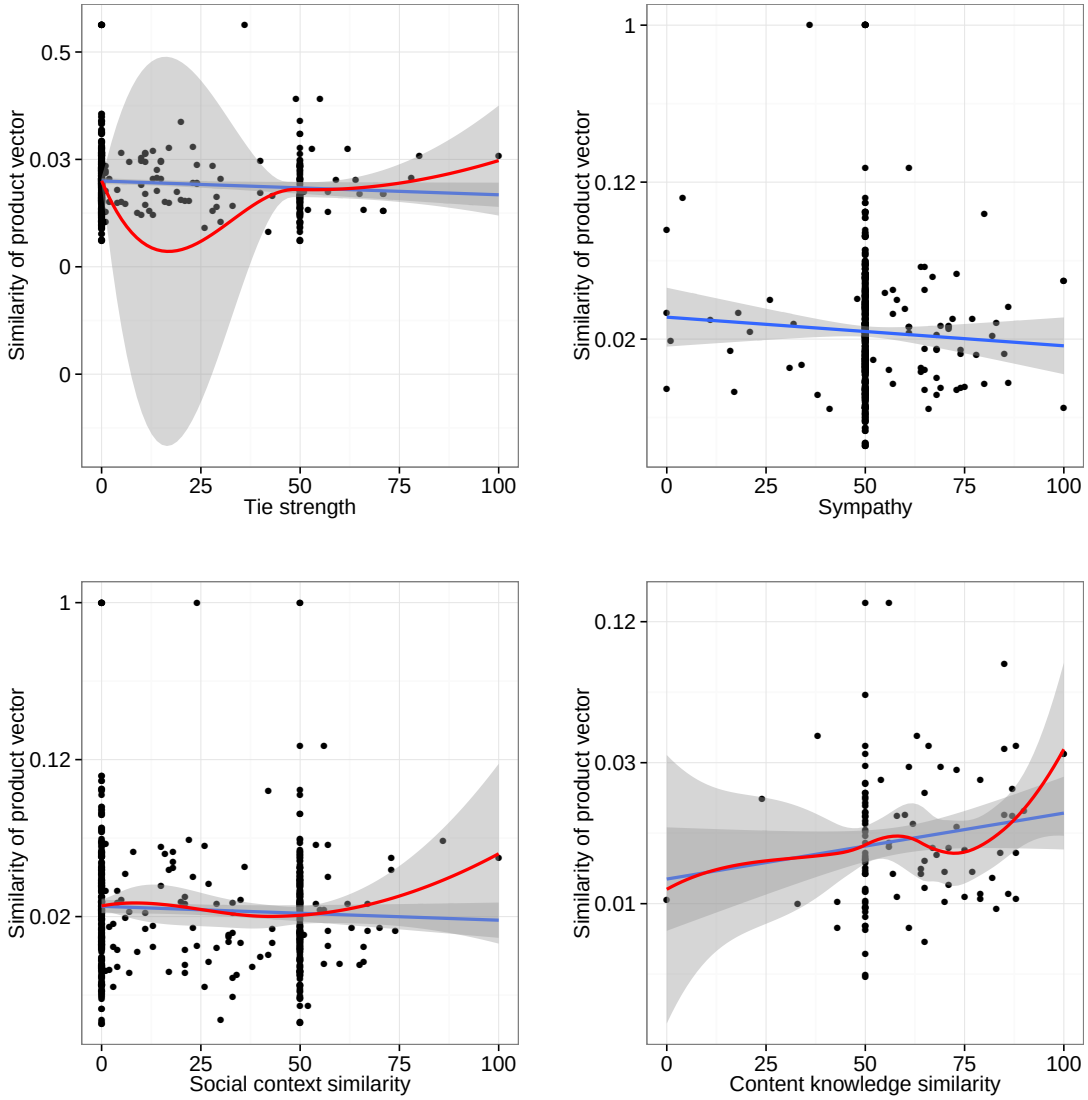


Figure 36: Correlation of social attributes and similarity of *bought products* on Amazon; plots show data for sim>0; blue line: linear regression, red line: LOESS regression (not available for sympathy); log-scale; gray area denotes 95% confidence interval

the following, these items will be referred to as "social items"). The group of other users to take the social items from (referred to as "friends") was selected following a randomly chosen strategy (out of eight available strategies: low/high tie strength, low/high sympathy, low/high social context similarity, low/high content knowledge similarity). The decision which item to add was based on the following criteria:

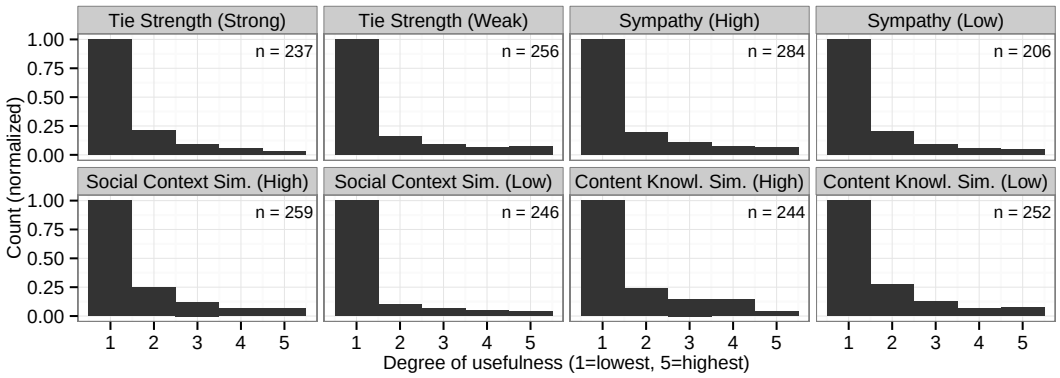
- The social item must have been bought or viewed by one of the "friends". While buying a product denotes a conscious and prudent decision, viewing a product requires much less effort and could be used more often (especially when considering that the group of participants consists of students with quite low incomes).
- The social item must be assigned to the same Amazon product category as one of the non-social result items obtained from Amazon.
- The social item must not already be part of the result items obtained from Amazon (to avoid double entries).
- Out of the remaining candidates, the one with the lowest distance to the query was chosen, measured by a modified version of the Levenshtein distance to allow fuzzy matching of substrings.[5]

For each item (social and non-social), the information how many people in the specific group have bought or viewed the respective item has been added. The participants did not know which strategy was used to determine "friends" (they did not even know that multiple strategies exist). Each item in the result list was rated by the user with regard to "usefulness" and "unexpectedness" using a 5-star scale (mapped to values 1-5, with 5 being best). In addition, for 50% of the result sets, one of the items retrieved from Amazon which has not been marked as bought or viewed from a friend already was presented in a way that the user would assume that a friend bought or looked at this item ("fake recommendation"). To summarize, the result set presented to the user consisted of the following items:

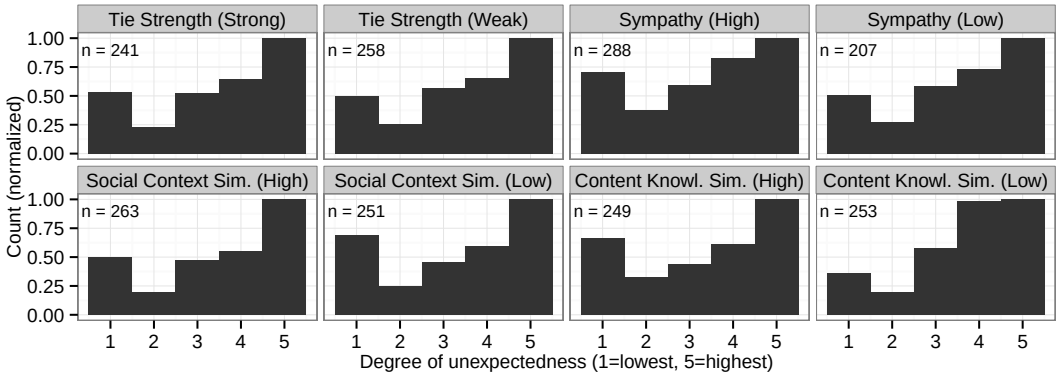
- Items obtained directly from Amazon, matching the query (3 items in total). A counter shows how many "friends" have bought or viewed each item. In 50% of the result sets, one of the items which has not been viewed or bought from the group of "friends" was marked as if it has been viewed or bought from "friends" ("fake recommendation").
- Items,
  - which are not part of the Amazon result list, but
  - which belong to a product category relevant for the query, and
  - which match the query using a function for fuzzy string matching, and which have been viewed or bought from one of the user's "friends",

  have also been added and marked accordingly (1 item for viewed, 1 for bought category; 2 items in total).

5 http://ginstrom.com/scribbles/2007/12/01/fuzzy-substring-matching-with-levenshtein-distance-in-python/ (retrieved 2016-01-10)



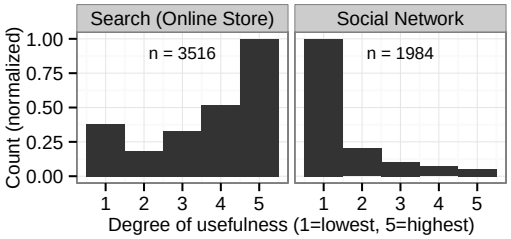
(a) Ratings of usefulness for items from other people, split by strategy to select items



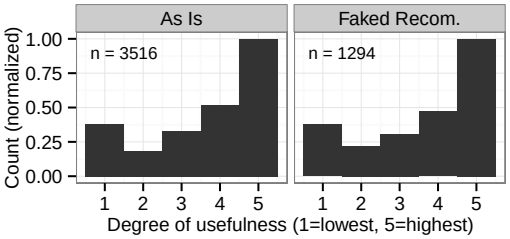
(b) Ratings of unexpectedness for items from other people, split by strategy to select items

Figure 37: Usefulness and degree of unexpectedness of "social items", split by strategy (Experiment 7c)

Figure 37a shows the usefulness ratings for the items added from friends, split by the eight strategies to identify friends. The plotted results suggest that there is no difference between the strategies. A Kruskal-Wallis test indicates a difference for at least one strategy (Kruskal-Wallis $\chi^2 = 20.942$, df = 7, p-value = 0.00). A Wilcoxon rank sum test testing the rating for usefulness between the high/low variants of each strategy pair reveals a statistically significant difference for the strategy pair based on social context similarity (W = 35,882, p=0.00). Given the limited differences between the strategies, a comparison of the ratings for usefulness between the result items obtained from Amazon and the ones added by social strategies is given in Figure 38a. The results show that the items obtained from Amazon are considered as highly useful in the majority of the cases, while the opposite is the case for the result items obtained from the "friends" group (Wilcoxon rank sum test, W = 5,878,400, p-value = 0.00). Figure 38b shows that the result items originally from Amazon and (wrongly) marked as being viewed/bought from "friends" do not differ in terms of usefulness from the other result items obtained from Amazon (Wilcoxon rank sum test, W = 2,284,900, p-value = 0.81).

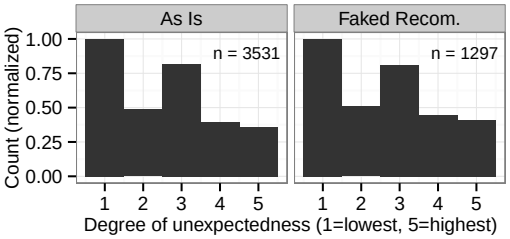


(a) Usefulness of result items obtained from Amazon and from social strategies



(b) Usefulness of result items obtained from Amazon, for "'normal" items and items allegedly bought/viewed by friends

(c) Degree of unexpectedness for result items obtained from Amazon and from social strategies



(d) Degree of unexpectedness for result items obtained from Amazon, for "'normal" items and items allegedly bought/viewed by friends

Figure 38: Usefulness and degree of unexpectedness for items from Amazon, social strategies, and "faked" social recommendations (Experiment 7c)

16.6.3.2 *RQ 2b: Does Social Context Imply a Valuable Contribution to Retrieving Information From the Unconscious Information Need (Serendipitous Information)?*

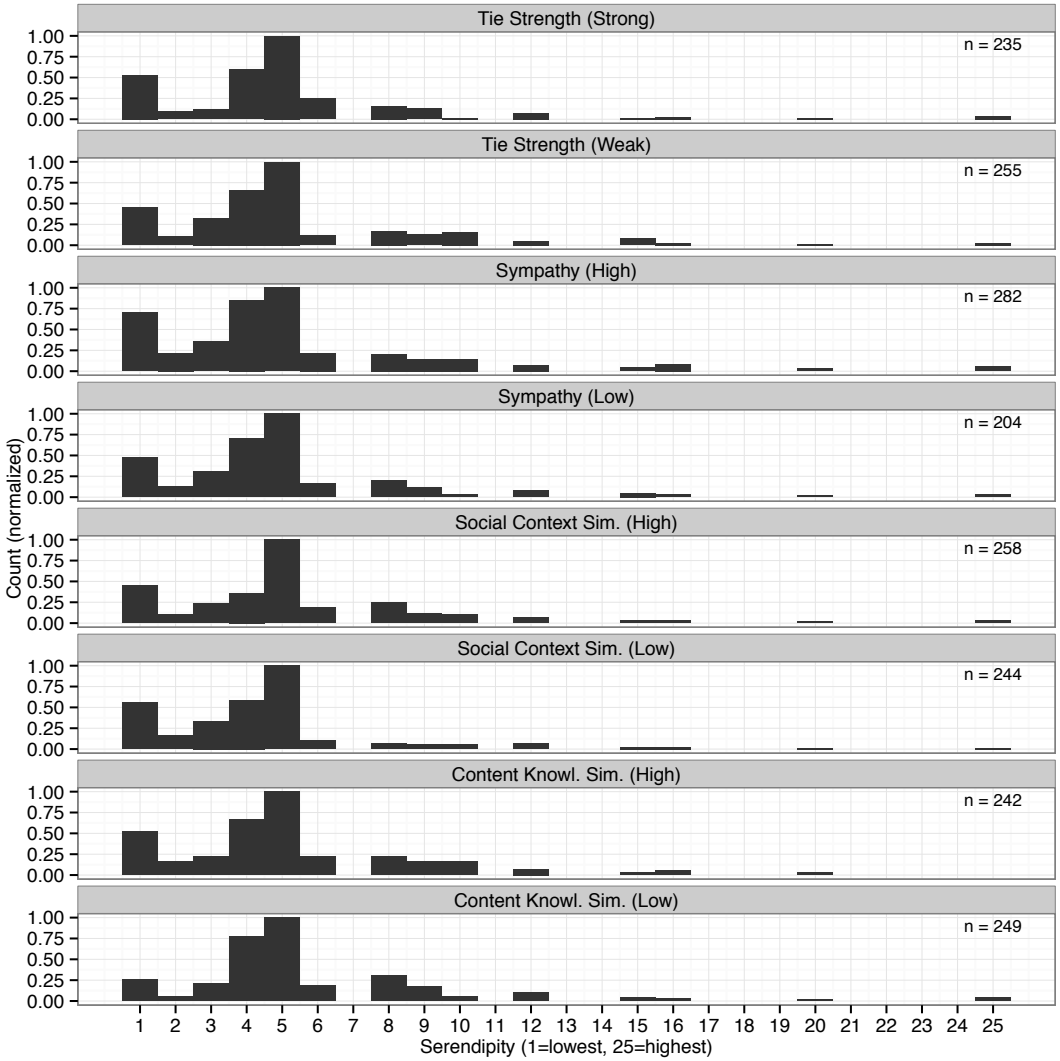
A serendipity value is calculated based on the product of the ratings for "usefulness" and "unexpectedness". Objective of the following analyses is to explore whether serendipity is linked to the strategy to select the provider of the additional result items or not. If serendipitous events happen significantly more often when using a certain strategy, this would suggest that the respective social attribute might be a useful predictor to add controlled serendipity to a result list.

Figure 39a suggests that the strategies to define the group of "friends" to take the result item from do not differ much, however, a Kruskal-Wallis tests indicates that at least one strategy leads to different serendipity values (Kruskal-Wallis $\chi^2 = 26.986$, df = 7, p-value = 0.00). Wilcox rank sum test confirms a statistically significant difference between the high and low social context similarity (p=0.00). Figure 39b illustrates the distribution of serendipity, split by the type of origin (directly from Amazon vs. taken from "friends"). The difference between both populations is confirmed using the Wilcox rank sum test (W = 4,807,700, p-value = 0.00).

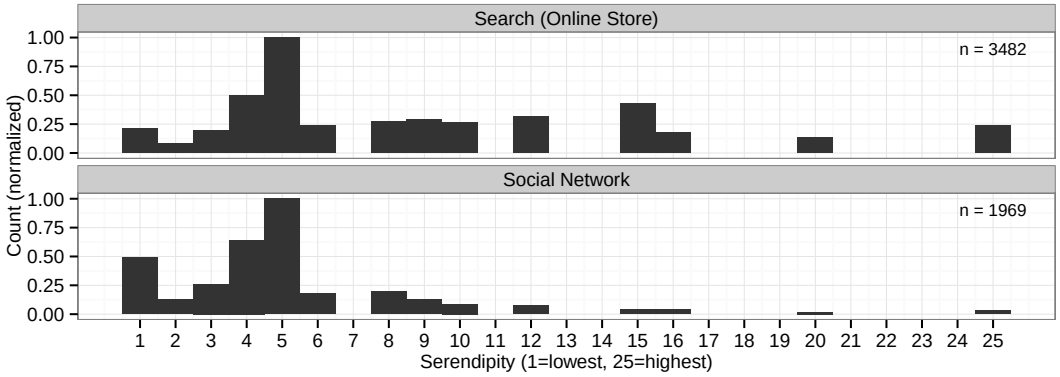
16.7 LIMITATIONS

In the design of Experiment 7a, each triple (query, information seeker, information provider) is interpreted as statistically independent measurement, since each instance only occurs once. The effects of multiple measurements have been considered in the random effects regression models, but have not been a primary concern in the experiment design. The objective of the experiment was to holistically test and generate hypotheses – we therefore decided to give the participants as much freedom as possible in defining queries and using the system. In a setting where variables are controlled to a higher degree, single effects could be isolated better: for instance, it would make sense to ask each information provider the same set of queries, sent from information seekers that fall into certain categories with regard to their social relation to the information provider. This approach would allow to build a model with the information provider as central component that depends on queries and the attributes of the social relationship to information seekers.

In addition, in Experiment 7b it is possible to increase the expressive power of the information space based on the users' browsing history: in the experiment, a limited number of topics (100) for the latent topic space was used. This parameter has been defined in advance based on experiments with a selected set of browsing histories. During these experiments, the complete browser history of two real-world user accounts has been downloaded and used to calculate different topic models with a variety of topics (5, 20, 50, 100, 200). The resulting topics have been visualized using word clouds. For both user accounts, the topic model with 100 topics led to a result which allowed the easiest interpretation. Due to operational necessities and privacy concerns, it was not possible to optimally adjust the number of topics for each user because the browsing histories' scope differed to a large extent among the participants. Approaches for a variable number of topics like (Wang et al., 2011) did not provide sufficiently stable results in the test cases we conducted. It would make sense to either get the complete browsing history for each participant to optimize the



(a) Serendipity values of items taken from "friends", split by strategy to define "friends"



(b) Serendipity values of items, split by origin

Figure 39: Distribution of serendipity, split by social strategies and result item origin (Experiment 7c)

parameters of the models or to elaborate on ways to define suitable parameters for a predefined set of situations and automatically detect the respective situation for each user's information space. Alternatively, it would make sense to merely consider the automatic mode as a supporting function for the information provider to reply to a received information need instead of expecting that the information seeker could get an automated reply directly.

In Experiment 7c, the results might have been influenced by the fact that the participants did not always fully distinguish "usefulness of the result item" and "relevance of the result item to the query". In the previously conducted experiments, the participants had to assess search results for relevance. This could have biased the participants (despite the difference was explicitly mentioned in the communication towards the participants). Furthermore, the way how "social" items got selected could have induced a bias. It could be the case that a product category system with a higher resolution or a different function to decide which item to inject in the result item list would have led to different results. A "social" item was only part of the result list if the item has not already been added by Amazon's search functionality itself, so the system added only those "social" items which would have not been added by Amazon. Following this approach allows us to only measure the cases where the "social" items are so different from the "normal" items that a common search engine used by Amazon is not capable of finding it – and increases the risk of including a non-fitting item. The social strategies therefore do not get the credit for identifying items which would have been identified by Amazon as well.

All datasets of Experiment 7 have in common that they have been created in a declared experimental situation. An experimental setup relying on behavioral data to a higher degree could possibly lead to different results (but would require to focus on isolated effects and could possibly introduce an interpretation bias, i.e. a bias caused by misinterpreting user activities).

Part IV

SUMMARY OF RESULTS AND DISCUSSION OF IMPLICATIONS

The following part consists of three chapters: in the first one (Chapter 17), the results of the experiments documented in Part III are briefly summarized for each Research Question. In the second chapter (Chapter 18), the derived implications for a Social Information Retrieval system are discussed. Chapter 19 closes this thesis with a critical perspective on the outcomes and an outlook on possible future research topics which could build upon or extend these results.

17

SUMMARY OF RESULTS

17.1 RESEARCH QUESTION 1: HOW DO SOCIAL CONTEXT AND INTERACTION ARCHETYPES INFLUENCE USERS' DATA SHARING SENSITIVITY IN VIEW OF SOCIAL INFORMATION RETRIEVAL APPROACHES?

In a social information retrieval scenario, two parties need to share information: the information seeker sends a query and thus releases information to the designated information providers who receive the query. The information provider replies with her response. Our findings suggest that the willingness to share data is impacted by social context and interaction archetype for both roles.

INFORMATION SEEKER The results of Experiment 2 (Chapter 11) suggest that information seekers feel much more comfortable to share their information needs with a precisely defined group of potential information providers. Forcing the information seekers to disclose their information needs to a wider audience (as it is done in classical SMQA approaches (Oeldorf-Hirsch et al., 2014)) causes hesitance to use social means of information retrieval. In Oeldorf-Hirsch et al.'s experiment, 76% of all information needs got routed to traditional search engines exclusively. In our study, only 42% of the information needs were exclusively sent to traditional search engines. The main difference between the two studies is that we offered an additional option to send the request to a set of carefully chosen friends directly. This option has been considered for 48% of all information needs (non-exclusively, Figure 7). Since the share of information needs routed to SMQA for both experiments is of equal size, we consider our results as quite reliable (Section 11.6), although our data is based on a survey, while Oeldorf-Hirsch et al. run a laboratory experiment (Section 11.7). The results confirm the hypothesis that information seekers consider sending queries in a social information retrieval scenario as a potential privacy threat and therefore need to be protected. People carefully maintain a certain "public persona" and do not want to put it at risk by asking potentially damaging questions. Possible ways to reduce this barrier to social information retrieval include an optional "anonymous mode" as explained in Section 7.2.2 and a way to specify the audience of queries on a precise level. In addition, Experiment 7a (Section 16.6.1.5) reveals that information seekers tend to select contacts with higher values for tie strength, sympathy, social context similarity, and content knowledge similarity (Figure 25). Tie strength appears to be the most important factor. This could be caused by the fact that the information seekers try to avoid huge social costs during the experiment (which could possibly occur when asking someone who is more distant).

INFORMATION PROVIDER The data collected during Experiment 7a (Section 16.6.1.1) suggests that information providers with higher values for tie strength and sympathy tend to reply to requests more often (Figure 18). An analysis of the reply's degree

of privacy and the social attributes tie strength, sympathy, social context similarity, and content knowledge similarity on the same data shows a mixed picture. While tie strength is suggested to have a significant positive impact on the reply's degree of privacy, sympathy's effect appears to be negative. As already stated earlier, tie strength's positive impact on sharing private information has been expected (Levin and Cross, 2004) and is confirmed by our findings, however, the negative effect of sympathy is surprising. One possible interpretation could be that sympathy does not necessarily need a close relationship, but can be a feature of the other person, while tie strength requires more closeness. Social context similarity and content knowledge similarity do not seem to heavily influence the privacy degree, since the variables are not statistically significant in any of the fitted models.

Experiment 7a also indicates that information providers' willingness to share information depends on the social interaction archetype: when receiving explicit requests, the willingness to share even private information is much higher than in other (proactive) sharing scenarios (Figure 23, Figure 22c). The results also suggest that the participants are reluctant to share information items with larger audiences, especially once a certain level of privacy is exceeded (Figure 22c). It is possible that information providers are willing to spend the additional effort to assemble a list of recipients (and do not fall back on the default "all friends" list) for information items with a degree of privacy higher than ~ 50 (where the line for LIMITED AUDIENCE crosses the line representing the larger group of FRIENDS in Figure 22c).

17.2 RESEARCH QUESTION 2: RELEVANCE AND SERENDIPITY OF RESULTS

17.2.1 *How Relevant Are Information Items Taken From Non-Public Information Spaces of Socially Close People When Satisfying Information Needs?*

Experiment 1 suggests that content created by socially close contacts is of higher relevance than content created by other people (Section 10.6.1). Experiment 7a (Section 16.6.1.3) reveals a statistically significant positive correlation of relevance with sympathy, tie strength, and social context similarity (Table 26). The effect is not fully confirmed in Experiment 7b (Section 16.6.2.1), where only a positive effect for tie strength is identified. To a certain extent this might have been caused by the setup of the information spaces (Section 16.7). Experiment 7a can be seen as being close to the "upper bound" of performance of a social information retrieval system with automatic routing and automatic query answering with topic based indices: since the matching and routing activities are done manually, no technical solution can negatively influence the results. On the contrary, Experiment 7b is a technical implementation and therefore potential technical shortcomings (e.g., with regard to LDA, Section 16.7) are reflected in the results. Experiment 7c (Section 16.6.3.1) confirms that socially close people have similar interests, i.e. look at the same products in an online shopping portal (and therefore can support each other in decision making, cf. Section 16.6.3.1). Having at least one product in common (i.e., "a common ground"), the correlation increases by a factor of ~ 4 (Figure 35). The effect can not be shown for bought products (Figure 36), which might be caused by the limited financial strengths

of the participants (students). In the following, each social attribute will be discussed in detail (based on the datasets from Experiments 7a, 7b, and 7c).

TIE STRENGTH Overall, Experiment 7a suggests a positive correlation of tie strength with relevance (Table 26). This effect is confirmed in Experiment 7b (high relevance increases probability for strong tie strategy, cf. Table 34). Tie strength appears to be positively correlated with the similarity of viewed products on Amazon (Pearson's r=0.354 / Spearman's rho=0.141 if at least one item has been viewed by both users, otherwise Pearson's r=0.078 / Spearman's rho=0.034, cf. Figure 35). While having first indications in the data that tie strength might increase the relevance of a search result, the effect observed in the data is quite small. With the products viewed on Amazon, the homophily effect (cf. (Tang et al., 2014)) was observed more clearly and demonstrated that the suitability of an information provider to answer an information need in a specific use case (e.g., information on products) could be higher for socially close information providers. For a social information retrieval system, considering more elaborated measures of tie strength (Gilbert and Karahalios, 2009) could show the effect on relevance more drastically.

SYMPATHY In Experiment 7a, sympathy was reported to have a positive correlation with relevance (Table 26). Experiment 7b does suggest an opposite effect which is not statistically significant (Table 34). Sympathy seems to positively influence the similarity of viewed products on Amazon in Experiment 7c (Pearson's r=0.388 / Spearman's rho=0.158 when focusing only on those observations where at least one product has been viewed by both users, otherwise Pearson's r=0.068/Spearman's rho=0.044, cf. Figure 35). Except for Experiment 7b, it has the same effect on relevance than tie strength discussed before, but appears to be smaller. For a social information retrieval system, explicitly considering sympathy might be more useful in the market model (cf. Section 7.1.3) to leverage existing goodwill instead optimizing relevance.

SOCIAL CONTEXT SIMILARITY The correlation coefficients in Experiment 7a suggest a slight positive correlation of social context similarity with relevance (Table 26), however, the other models are statistically not significant. In Experiment 7c, social context similarity seems to have a positive impact on the similarity of viewed products on Amazon (Pearson's r=0.317/Spearman's rho=0.179 when focusing only on those observations where at least one product has been viewed by both users, otherwise Pearson's r=0.054/Spearman's rho=0.020, cf. Figure 35). A convincing reason to consider social context similarity (as measured in the experiment) in a social information retrieval system is not given in the data. While a positive effect on relevance could be plausible due to a common environment and norms, it is possible that the participants did not show sufficient variance in the data to reveal such a relation (e.g., all shared a common social context, the university and the major subject).

CONTENT KNOWLEDGE SIMILARITY In models fitted on the datasets of Experiment 7a and 7b, content knowledge similarity does not influence relevance on a statistically significant level (Table 26). In Experiment 7c, similarity in content knowledge slightly correlates with similarity of viewed Amazon products (Pearson's r=0.063,

Spearman's rho=0.087, cf. Figure 35). An interesting observation for content knowledge similarity is that having at least one viewed Amazon product in common does not increase the correlation results (unlike as for the other social attributes). Thus, the data does not reveal sufficient evidence to consider content knowledge similarity in the routing process of a social information retrieval system to maximize relevance.

17.2.2 *Does Social Closeness Imply a Valuable Contribution to Retrieving Information From the Unconscious Information Need (Serendipitous Information)?*

Experiments 7a (Table 29), 7b (Section 16.6.2.2), and 7c (Section 16.6.3.2) did not confirm that tie strength or sympathy have any impact on serendipity. Instead, content knowledge similarity (Experiment 7a and 7b) and social context similarity (Experiment 7c) have been identified to positively correlate with serendipity. The results suggest that serendipity requires a certain overlap of content knowledge between the information seeker and the information provider in order to evolve – without this common ground, the background knowledge which is necessary to transport, understand, and interpret the information need and the response might not be sufficient in order to significantly produce serendipity. As a rather extreme fictional example, a biologist who has a problem interpreting the data of an experiment could ask a computer scientist for help. The biologist first needs to be able to explain the problem in a way that the computer scientist understands the problem to a sufficient degree, so that she can apply her methodological expertise on the problem. In a proceeding step, the computer scientist has to explain the reply and ensure that that the biologist can transfer it to her knowledge domain. A higher overlap of content knowledge therefore increases the chances for a successful exchange of knowledge, while a higher diversity of content knowledge could foster serendipitous effects. Since increasing serendipity by asking socially close contacts was one of the initial hypotheses to motivate social information retrieval, this finding could be interpreted as a limit of social information retrieval: asking contacts with a higher overlap in content knowledge might benefit from social closeness (e.g., to receive an answer), but does not inherently rely on social ties.

17.3 RESEARCH QUESTION 3: WHICH SOCIAL CONCEPTS INFLUENCE THE USERS' ROUTING DECISIONS?

Experiment 7a (Section 16.6.1.5) confirms that information seekers select socially close contacts as designated information providers (Figure 25). A logistic regression model (Table 30) reveals that tie strength is the factor with the highest effect. Sympathy, content knowledge similarity, and social context similarity also have positive effects on the odds and are statistically significant. The findings are confirmed by a logistic regression model with random effects to account for multiple occurrences of information seekers (Table 31). The majority of the selected information providers did not report to have the impression that they have to reply to a query because of differences in social authority ranks, i.e. the values for Fiske's AR component (Section 2.4.2, (Fiske, 1992)) were not high. The selected social edges exhibit a strong CS component, while the EM/MP component was rather weak (Figure 26). The information

providers confirmed the information seekers' perception of the edges' components (Figure 27). Information seekers chose the information providers mostly based on their content or local knowledge, social closeness, or taste similarity (Figure 28).

17.4 RESEARCH QUESTION 4: WHICH CATEGORIES OF INFORMATION NEEDS COULD BENEFIT FROM SOCIAL INFORMATION RETRIEVAL?

According to the results of Experiment 2 (Section 11.6), information seekers prefer to route information needs with a high interaction factor (Favor, Invitation, Offer, Social Connection, Rhetorical, Recommendation, Opinion) to "social" targets (Figure 6). Experiment 3 (Section 12.6.2) suggests that information items that do not require formal expertise to create also show a high degree of "sociality". Surprisingly, content from FACTUAL KNOWLEDGE & NEWS shows high correlation with the attributes of "sociality". This might be caused by the fact that people are interested in the interpretation of current events or facts (Table 10) and supports the social identity theory (Tajfel and Turner, 1979) (people need to know the prevalent position on common topics within their perceived ingroup). With the dataset collected during Experiment 7a (Section 16.6.1.6), we could not show statistically significant performance differences between the content or type categories. Being able to show that certain types of queries yield better performance than others would have helped to identify additional beneficial use cases for social information retrieval and to adapt the system further to the information seekers' needs.

18

IMPLICATIONS FOR SOCIAL INFORMATION RETRIEVAL SYSTEMS

In the following, implications of our findings for social information retrieval systems are discussed.

PRIVATE INFORMATION FROM SOCIALLY CLOSE PEOPLE IS RELEVANT In general, our results suggest that information from within one's own social network is considered useful by the participants – especially for more subjective, individual queries, it can be seen as a valuable add-on to the traditional search engines which follow the library paradigm (Chapter 1). This can be interpreted as a confirmation that social information retrieval systems constitute a useful object of research (cf. Chapter 10, Chapter 16). Our findings in a social information retrieval scenario with manual routing and manually entered replies (Experiment 7a, Section 16.6.1.3) confirmed that tie strength and sympathy positively correlate with relevance. In an additional experiment, where the selection of information providers and the composition of replies have been done automatically by the system (Experiment 7b, Section 16.6.2.1), tie strength was also identified as positive factor. Since the automatic scenario in Experiment 7b introduced additional influencing variables to the experiment (implementation of the information space, quality of matching algorithm, automatic routing function), we cannot exclude that sympathy might be relevant.

A REACTIVE APPROACH FOSTERS INFORMATION SHARING BEHAVIOR The majority of existing approaches that try to make use of the information and expertise of one's social network often rely on proactively shared information. Our findings confirm that information providers are reluctant to share information without a specific reason in advance (cf. Section 16.6.1.2). Following a "need to know" regime and sharing only explicitly requested information circumvents information providers' resistance to publish information. A social information retrieval system therefore should rely on the reactive interaction model, where information seekers query information providers (and should not assume that information providers already shared enough information in advance).

NAIVE INFORMATION RETRIEVAL TECHNIQUES MIGHT BE SUFFICIENT TO ORGANIZE INFORMATION SPACES The results of Experiments 5 (Chapter 14) and 6 (Chapter 15) provide hints that established IR approaches like TF (individually or enhanced with a local IDF component) might be sufficient to organize the information providers' information spaces. On the datasets used in our experiments, TF did not perform much worse than more complex approaches like LDA or ESA. Using a less complicated algorithm like TF (enhanced with a local IDF measure) would reduce complexity significantly and therefore would lower the required computing power of the devices that maintain the information spaces. To avoid any wrongly shared in-

formation items, the information items identified as relevant should be presented to the information provider for final approval before being shared with the information seeker. In a more restricted setup, a heuristic like a trained classifier could be put in place to reduce response times. However, since a single wrongly classified item could have severe social consequences, the heuristic should be either very conservative or the intended use case should bear little potential risk (e.g., in a professional work environment with a reduced information space).

KNOWLEDGE ADVERTISEMENT MIGHT REQUIRE MORE COMPLEX METHODS To ensure a meaningful routing process, users must have an idea of the available knowledge in their social network. Human social relationships are manifold and so are the preferences to communicate a certain set of expertise. Following the same conservative approach for expertise advertising as for sharing items from the information space, an information provider needs to explicitly state which forms of expertise she wants to advocate towards a user. This knowledge profile vector has to fulfill two opposing requirements: it has to be comprehensive enough to allow an information routing decision, while at the same time it must have a limited number of dimensions to allow a manual approval by the respective information provider, whose expertise is modeled in the vector. In the conducted experiment (Experiment 4, Chapter 13), approaches based on LDA and ESA performed better than TF-IDF. TF-IDF would be no valid option due to the large number of dimensions of the vector space. ESA and LDA seem to be better suited, since e.g. the network structure of Wikipedia would allow to rely on the links among concepts to use clustering mechanisms to reduce the dimensionality of the knowledge profile vector. LDA could be run with a manageable number of topics as well.

To improve the user's convenience and the performance of the system, a recommendation system should suggest suitable contacts, based on their assumed knowledge, the social costs, and the expected serendipity of the result (taking content knowledge similarity into account, cf. Experiment 7a and 7b in Chapter 16). The exact estimation of the parameters for such a system needs to be evaluated separately.

INFORMATION SEEKER'S PRIVACY (ALSO) NEEDS PROTECTION When thinking about privacy protection in social search, the first person that comes to mind is the information provider. With the reactive sharing approach and the rather conservative knowledge advertisement process, we addressed the information provider's requirements for privacy to a substantial extent. However, when using a social search system, the information seeker's visibility changes: while she can stay anonymous in traditional search engines, social search engines need to explicitly offer such a feature. The results of Experiment 2 (Chapter 11) suggest that allowing the information seeker to define the audience of an information need is a first step to increase the acceptance of such a system. Furthermore, an anonymous mode as detailed in the proposed concept (Section 7.2.2) might reduce the information seeker's reluctance even further.

INFORMATION SEEKERS PREFERABLY USE SOCIAL INFORMATION RETRIEVAL FOR CERTAIN TYPES OF QUERIES Information seekers would intuitively send queries covering recommendations, opinions, factual knowledge, and personal experience to

social information retrieval systems (Experiment 7a, Section 16.6.1.6). We could not find enough evidence to show that these categories also perform better than others in terms of relevance or serendipity metrics. Nevertheless, since information seekers most likely feel comfortable to share those types of information needs with their friends, the system could be adjusted to make these kind of queries as easy as possible (e.g., by semantically representing items to comment on, cf. next paragraph).

SEMANTICALLY RICH INFORMATION OFFERS ADDITIONAL BENEFITS As Experiment 7c (Section 16.6.3) revealed, socially close people are also interested in similar products, so the product/experience domain seems to be a reasonable use case for social means to satisfy information needs. Thus, considering social data retrieval next to social information retrieval to not only rely on unstructured data, but allow the exchange of semantically rich information like products of interest or URLs of visited websites appears to be a promising approach.

PARTICIPANTS SEEM TO ACT RATHER ALTRUISTIC The participants of Experiment 7 (Section 16.6.1) acted surprisingly altruistic. In terms of Fiske's elementary forms of social interaction (cf. Section 2.4.2), the components of EM/MP and AR have been rather low, indicating that the information providers helped the information seekers without explicitly reporting that they would expect something in return or feel forced by differences with regard to social rank / authority. In addition, the information seekers did not feel obliged to compensate the received favors. It is unclear whether this behavior can be ascribed to the sample group (mostly students), the experiment itself or a bias caused by the participants' preventing them to answer honestly (e.g., because no one dared to explicitly state that helping someone might be a reciprocal action). Therefore, it is difficult to generalize our findings without any further investigation.

19

CONCLUSION

Following the village paradigm introduced in (Horowitz and Kamvar, 2010), the central part of the present thesis is the evaluation of chances and limits of social information retrieval systems from technical and social perspectives, on the basis of a sufficiently generic but also specific concept (Part II) of a social information retrieval system. The technical evaluation is important to confirm that the approach can be considered as a useful enhancement to the prevailing library approach to satisfy information needs. The social evaluation is necessary to understand how a user would work with such a system and to adjust it accordingly. Searching for information unveils details about all actively involved parties of the search process and therefore needs to consider human's social interaction traits. The main objective of the research project was to initially gain a first comprehensive understanding of the problem domain. Therefore, multiple topics have been in focus and led to various experiments:

- Experiment 1 (Chapter 10) lay the foundation for the social information retrieval approach, giving evidence that content from socially close people is of higher relevance than content from others.
- Experiment 2 (Chapter 11) confirmed that information seekers still consider their social network as valuable source of information, despite having access to traditional search engines operating on tremendous data collections.
- Experiment 3 (Chapter 12) and Experiment 7a (Section 16.5.2) suggested that users would most likely prefer social information retrieval for specific types of information needs (e.g., recommendations and opinions based on personal experience).
- Experiment 4 (Chapter 13) evaluated the routing mechanisms of the proposed concept and showed that LDA and ESA would be suitable approaches on the tested dataset.
- Experiment 5 and 6 (Chapter 14, Chapter 15) compared several implementation variants to represent information in information spaces. The datasets used for evaluation have been obtained from the Stackexchange communities[1]. More advanced approaches like ESA or LDA could not demonstrate advantages in comparison to well-established methods like TF-IDF on the tested datasets.
- Experiment 7 (Chapter 16) consisted of three large-scale social information retrieval scenarios with >100 participants. The first one (Experiment 7a, Section 16.5.2) represented a manual scenario, where the information provider was determined manually by the information seeker. Furthermore, the information provider replied manually following a predefined process. The main objective

1 `http://www.stackexchange.com` (retrieved 2016-01-17)

was to get insights about the social dynamics and to estimate an upper bound for the performance of a social information retrieval system. Experiment 7b (Section 16.5.3) was a revision of Experiment 7a, including an automated routing and answering mechanism. Both experiments suggested that social attributes like tie strength and relevance are correlated; however, for Experiment 7b the impact of additional influencing parameters caused by the technical implementation must be considered (cf. Section 16.7). Experiment 7c (Section 16.5.4) simulated an explicit semantic social information retrieval scenario for product search. The results revealed that social closeness and the number of jointly viewed products correlate. This effect could be interpreted as a confirmation that the users within one's social network are a competent source for product recommendations (and at the same time have similar valuation preferences).

In the following, the responses to the research questions introduced in Chapter 1 are briefly summarized. For a more in-depth discussion, please refer to Chapter 17.

- Research Question 1 (*"How do social context and interaction archetypes influence users' data sharing sensitivity in view of social information retrieval approaches?"*): Information seekers and information providers are willing to share more information (queries, responses) with socially close people. Information providers tend to share more (and more sensitive) information in a reactive interaction model where data is requested (and not expected to be shared in advance). Information seekers prefer to precisely define the audience of queries.

- Research Question 2a (*How relevant are information items taken from non-public information spaces of socially close people when satisfying information needs?*): The results of both evaluated scenarios (Experiment 7a, 7b, Section 16.6.1, Section 16.6.2) suggested that social closeness (specified as "tie strength") is positively correlated to relevance.

- Research Question 2b (*Does social context imply a valuable contribution to retrieving information from the unconscious information need (serendipitous information)?*): In both scenarios (Experiment 7a, 7b, Section 16.6.1, Section 16.6.2), social closeness did not correlate with measures for serendipity. Instead, content knowledge similarity was identified as positive correlation. Considering the results from other studies like (Kukla et al., 2012) which suggest that people reply to requests from socially close information seekers with a higher probability, higher tie strength could nevertheless help to receive an answer from someone with a high degree of content knowledge similarity. Even if our results do not suggest a direct relation, an indirect effect is still possible and subject of future research.

- Research Question 3 (*Which social concepts influence the users' routing decisions?*): The social attribute with the largest effect on a user's probability to get selected as information provider was tie strength. All four social attributes (tie strength, sympathy, social context similarity, content knowledge similarity) were positively correlated with the probability of getting selected as information provider.

- Research Question 4 (*Which categories of information needs could benefit from social information retrieval?*): Individual and interactive types of information needs

(e.g., recommendations, opinions, sharing of personal experience) and information needs from content areas where limited professional expertise is required are suggested candidates for social information retrieval. However, the collected datasets do not allow to infer that these categories of information needs perform substantially better than others.

As a first iteration of the design science cycle, this attempt helped to build a reference system for social information retrieval that can be used to collect additional insights about the technical functionality and the determining social traits of the users. The experiments suggest that social information retrieval is a promising approach, but is most likely not used as only means of information retrieval. While our findings suggest that social closeness is considered helpful when supporting each other, a social search system can only fully demonstrate its strengths in suitable use cases: for "standard queries" like navigational or factual information needs, traditional search engines will most likely outperform other approaches in terms of response time, variety, and even quality of answers. Whenever a more personal, individualized information need should be satisfied, concepts of social homophily or trust could allow the social information retrieval system to support the user with a more personal response than traditional search engines. Focusing on specific scenarios like product search, the privacy-preserving aggregation of items could illustrate the potential of the idea quite fast.

For future work, the presented concept could be considered as framework and isolated topics and challenges could be treated with a greater focus. It is quite likely that the participants' willingness to share information depends on individual social traits (e.g. modeled via OCEAN (Goldberg, 1993)), which have been considered implicitly in the random effects models (cf. Section 16.6.1.1). Confirming this assumption would possibly allow to enhance a classifier which helps the information provider to decide whether or not to share a certain information item that has been identified as relevant for an information seeker's query. The experiments have been conducted to either confirm fundamental hypotheses or to evaluate parts of the social information retrieval concept. The components of the concept have been implemented in a way to allow experimental testing – an evaluation in a more "natural" environment could provide additional insights. Based on the results of Experiments 5 and 6 (Chapter 14, Chapter 15), the implementation of a mobile prototype could possibly be easier as initially presumed, since TF and TF-IDF did perform sufficiently well on the tested collections. This would also address the topic sparsity problem faced in Experiment 7b (Section 16.7). The automatic mode evaluated in Experiment 7b could also be seen as an extension of the manual mode (Experiment 7a) to only support the information provider (like a local search engine). The results from Experiment 5 (Chapter 14) and 6 (Chapter 15) are backed by various large datasets, however, the datasets always covered a delimited content area (e.g., travel, Islam, etc.) and therefore the algorithms might perform differently on a more diverse dataset. In addition, possible parameter adjustments might change the results as well (cf. Section 14.8). The participants consisted of students, their individual social network, and participants in online surveys. Apart from the experiments conducted with existing data (Experiment 1, 4, 5, 6), all participants have been aware that they were part of an experimental study. This could have influenced their judgment and behavior. The social capital model (Section 7.1.3)

has only been interpreted based on the ratings of Fiske's elementary forms of sociality. With a fully implemented prototype, a research approach following the principles of behavioral science could lead to more expressive results than the survey approach used in Experiment 7 (Chapter 16).

The thesis performs a first iteration in the design science cycle towards a social information retrieval system following the village paradigm (cf. Chapter 1) and hopefully contributes to the improvement of one of the most important tasks in the information society: finding a right answer to a question.

A

APPENDIX

A.1 EXPERIMENT 3

A.2 EXPERIMENT 7

A.2.1 *Variables*

Variable	Scale type	Question
Real name	nominal	"Real name" (registration form)
Matriculation number	nominal	"Matriculation number" (registration form)
Username	nominal	"Username" (registration form)
Photo	/	"Profile photo" (registration form)
Email address	nominal	"Email address" (registration form)
Password	nominal	"Password" (registration form)
Gender	nominal	"Gender" (registration form)
GPS home coordinates	interval	"Please enter the coordinates of your main place of residence (lat./long.)" (incl. a link to a website with detailed explanations on how to get the coordinates using Google Maps)
Amazon language setting	nominal	"What is your Amazon language setting?"
Facebook network	/	(extracted via SNExtractor, cf. Section A.2.2)

Table 37: Variables gathered during preparation phase of Experiment 7 describing each participant

CATEGORY / WEBSITE	CATEGORY / WEBSITE
AUTOMOTIVE	HOME & GARDEN
fordforums.com	gardenweb.com
honda.com	groupon.com
teslamotors.com	yelp.com
BUSINESS & PROFESSION	LIFESTYLES
alibaba.com	ebay.com
monster.com	netflix.com
paypal.com	totalbeauty.com
ENTERTAINMENT	REAL ESTATE
9gag.com	movoto.com
imgur.com	realtor.com
youtube.com	zillow.com
ETHICS & PHILOSOPHY	RECREATION
scu.edu	classic.party4.com
tufts.edu	sailinganarchy.com
FACTUAL KNOWLEDGE & NEWS	summitpost.org
broadwayworld.com	SOCIETY
news.yahoo.com	answers.yahoo.com
reddit.com	facebook.com
FINANCE & INSURANCE	stackoverflow.com
finance.yahoo.com	SPORTS
money.cnn.com	espn.cricinfo.com
usaa.com	espn.go.com
FOOD & DRINK	sports.yahoo.com
allrecipes.com	TECHNOLOGY
food.com	engadget.com
liquor.com	gizmodo.com
GAMES	thinkgeek.com
eu.battle.net	TRAVEL
ign.com	booking.com
twitch.tv	tripadvisor.com
HEALTH	TRIVIA
mayoclinic.com	funtrivia.com
nia.nih.com	jetpunk.com
webmd.com	mentalfloss.com

Table 36: Websites used in Exp. 3, based on Alexa's categories and top sites (`http://www.alexa.com/topsites/category` (retrieved 2015-02-04)), with slight adjustments to reflect the content of the exemplary URLs

Variable	Owner	Question
Query	IS	"Please choose a query which reflects an information need you already have and which is suited (in your opinion) to get answered by your friends / social contacts."
IP	IS	"Provide the hash of the person you would like to add as a recipient to the query", incl. details on how to use SNTranslator (Section A.2.2) to get the hash value
Reason for IP	IS	"Why do you think that this recipient is a good choice?"
Reply	IP	"Your reply to the question and message to the question asker (please fill your answer in the text field or send it directly to the sender, e.g. via email, whatsapp, etc.). It is essential that you only use information you already know (e.g. URLs for websites you visited in the past, details you read/heard about, etc.) - please do not start a fresh search at Google/any other search engine and come up with something new you did not know before (however, it is okay to search for a specific resource in order to rediscover it)"
Reply done	IP	"Did you answer the request?"
Reason reply not done	IP	"Why did you not reply to the request (optional, unless you specified above that you did not reply to the query)"
Already shared information	IP	"Did you already share the information in the reply to your friend somewhere in the social media sphere (social networking platform, social bookmarking service, weblog, etc.)?" - options for answer (ordinal scale): No; yes, but very limited audience; yes, friends only; yes, publicly
Reason already shared information	IP	"Why did you share / did you not share the information on social media before?"
Would share information	IP	"Would you share the information you gave to the requester on a social networking platform like Facebook?" - options for answer (ordinal scale): No; yes, but very limited audience; yes, friends only; yes, publicly
Reason would share information	IP	"Why would/wouldn't you share the information on a social media platform like Facebook?"
Would share explicitly asked	IP	"Would you share the information with a friend if she/he explicitly asks you for help?" - options for answer: Yes; no
Reason would share explicitly asked	IP	"Why would/wouldn't you share this information in this case?"
Privacy of information	IP	"How "private" do you consider the information in your reply?" - (ratio) scale from 0 ("Not private (could get published in a newspaper)") to 100 ("Highly confidential") using slider, default: 50
Communal sharing	IP	"Would you help the person who asked the question no matter what, i.e. in any situation?" - (ratio) scale from 0 ("No") to 100 ("Yes") using slider, default: 50
Authority ranking	IP	"To which degree did you help the question asker because of differences in social rank or status (e.g. boss vs. staff, caring parent vs. child)?" - (ratio) scale from 0 ("Low") to 100 ("High") using slider, default: 50
Equality matching / market pricing	IP	"Do you think that the person you helped 'owes' you something because you did her/him a favor by answering your question?" - (ratio) scale from 0 ("No") to 100 ("Yes") using slider, default: 50

Tie strength	IP	"How strong is your relationship to the question asker?" - (ratio) scale from 0 ("Weak") to 100 ("Strong") using slider, default: 50
Content knowledge similarity	IP	"How similar is this person to you in terms of content knowledge?" - (ratio) scale from 0 ("Very different") to 100 ("Very similar") using slider, default: 50, optional checkbox: "I do not know whether this person's content knowledge is similar to mine or not"'
Social context similarity	IP	"How similar are your social contexts?" - (ratio) scale from 0 ("Very different") to 100 ("Very similar") using slider
Sympathy	IP	"How much sympathy do you have for the other person? (all information you provide will stay private, i.e. no other user will be able to see it)" - (ratio) scale from 0 ("Not likeable at all") to 100 ("Highly likeable") using slider, default: 50
Satisfaction of information need	IS	"How much did the information provided by this person help to satisfy your information need?" - (ratio) scale from 0 ("Did not help at all") to 100 ("Information need fully satisfied") using slider, default: 50
Relevance of reply	IS	"Is the reply you received relevant for your query?" - (ratio) scale from 0 ("Not relevant at all") to 100 ("Highly relevant") using slider, default: 50
Unexpectedness of reply	IS	"Did the reply contain information you did not expect or that was not obvious; did the content surprise you?" - (ratio) scale from 0 ("Reply did not surprise me") to 100 ("Reply was highly unexpected") using slider, default: 50
Personalization of reply	IS	"To which degree was the information you received personalized to you as a specific person? (e.g. did the information provider include personal knowledge about you in the reply)" - (ratio) scale from 0 ("Not personalized at all") to 100 ("Highly personalized") using slider, default: 50
Communal sharing	IS	"Do you think that the person you asked for help will help you no matter what, i.e. in any situation?" - (ratio) scale from 0 ("No") to 100 ("Yes") using slider, default: 50
Authority ranking	IS	"To which degree do you think that the person you asked for help helps you because of differences in social rank or status (e.g. boss vs. staff, caring parent vs. child)?" - (ratio) scale from 0 ("Low") to 100 ("High") using slider, default: 50
Equality matching / Market pricing	IS	"Do you think that you owe something to the person you asked the query because she did you a favor by answering your question?" - (ratio) scale from 0 ("No") to 100 ("Yes") using slider, default: 50
Additional potential IPs	IS	"Please enter the hashed names (using SNTranslator tool) of up to three other people who would have also been good (or maybe better) contacts to ask this query"
Additional potential IPs / reasons	IS	Reason why this would have been a good contact to solve the query and reasons why you did not choose this contact

Table 38: Variables describing the manual query approach in Experiment 7a

Variable	Owner	Question
Reason forward request	IP^n	"Why do you want to forward this request?"
Reason for IP^{n+1}	IP^n	"Why do you want to forward the request to the person you have in mind?"
Communal sharing $IP^{n,IS}$	IP^n	"Would you help the question asker no matter what, i.e. in any situation?"
Authority rank $IP^{n,IS}$	IP^n	"To which degree do you help the question asker because of differences in social rank or status (e.g. boss vs. staff, caring parent vs. child)?"
Equality matching / Market pricing $IP^{n,IS}$	IP^n	"Do you think that the question asker owes you something because you did her/him a favor by forwarding this question?"
Communal sharing $IP^{n,n+1}$	IP^n	"Do you think that the person you forward the question to will help you no matter what, i.e. in any situation?"
Authority rank $IP^{n,n+1}$	IP^n	"To which degree do you think that the person you forward the question to helps you because of differences in social rank or status (e.g. boss vs. staff, caring parent vs. child)?"
Equality matching / Market pricing $IP^{n,n+1}$	IP^n	"Do you think that you owe something to the person you forward the query to because you bother her/him with this question?"
Tie strength $IP^{n,n+1}$	IP^n	"How strong is your relationship to the person you forward the query to?"
Content knowledge similarity $IP^{n,n+1}$	IP^n	"How similar is this person to you in terms of content knowledge?"
Social context similarity $IP^{n,n+1}$	IP^n	"How similar are your social contexts?"
Sympathy $IP^{n,n+1}$	IP^n	"How much sympathy do you have for the person you forward the question to? (all information you provide will stay private, i.e. no other user will be able to see it)"

Table 39: Additional variables describing the manual query process when queries get forwarded in Experiment 7a

A.2.2 *Technical Architecture*

The experiment was conducted using a central web application in combination with various add-on programs executed on the participants' computers. Table 40 gives an overview of the main components and explains their roles.

Experiment	Name	Purpose
7a, 7b, 7c	Web application	Upload data, fill out surveys, submit, and reply to queries
7a, 7b, 7c	SNExtractor	Assess the edge attributes of the individual Facebook friends (tie strength, sympathy, social context similarity, content knowledge similarity)
7a	SNTranslator	Convert the plain text name of a friend to a hash value (used to anonymize the data)
7b	generate_whitelist	Create a list of visited domains based on the participant's web browsing history used as a whitelist in the subsequent processing steps
7b	history_crawler	Download the URLs taken from the participant's web browsing history when listed in the whitelist
7b	indexer	Calculate topic models (LDA, (Blei et al., 2003)) based on the downloaded websites
7b	URLIdentifier	Transform index positions used in the topic models back to URL strings, used to preserve the participant's privacy
7c	AmazonViewedProductsToFile	Analyze the participant's browsing history and extract viewed Amazon products to a text file
7c	AmazonParser.js	Extract all products bought from Amazon website
7c	AmazonBoughtProductsToFile	Convert the output of AmazonParser.js to a text file

Table 40: Technical components for the social information retrieval experiment (Experiment 7)

In the following paragraphs, the components will be introduced briefly.

WEB APPLICATION The web application forms the central element of the experiment: It helps to coordinate the participants' tasks (registration, upload of data, etc.), offers the platform for manual and automatic query modes, and builds the interface for the social product search experiment. It also offers an HTTP interface for SNExtractor to extract the list of participants.

SNEXTRACTOR To be able to learn from the participant's actions (e.g., which contacts are consulted for which type of question), it is important to get an idea about her social network. During the experiment, the participant's Facebook friends and the other participants of the experiment build the participant's "social network". SNExtractor is implemented in Python, runs on the participant's computer, reads both lists of friends, and asks the participant to assess tie strength, content knowledge similarity, social context similarity, and sympathy for each participant and for at least 50 Facebook friends. Friendship relations from Facebook are extracted using Facebook's backup feature which allows to download a copy of the complete Facebook profile. SNExtractor generates two output files. Both files use a format that is originally

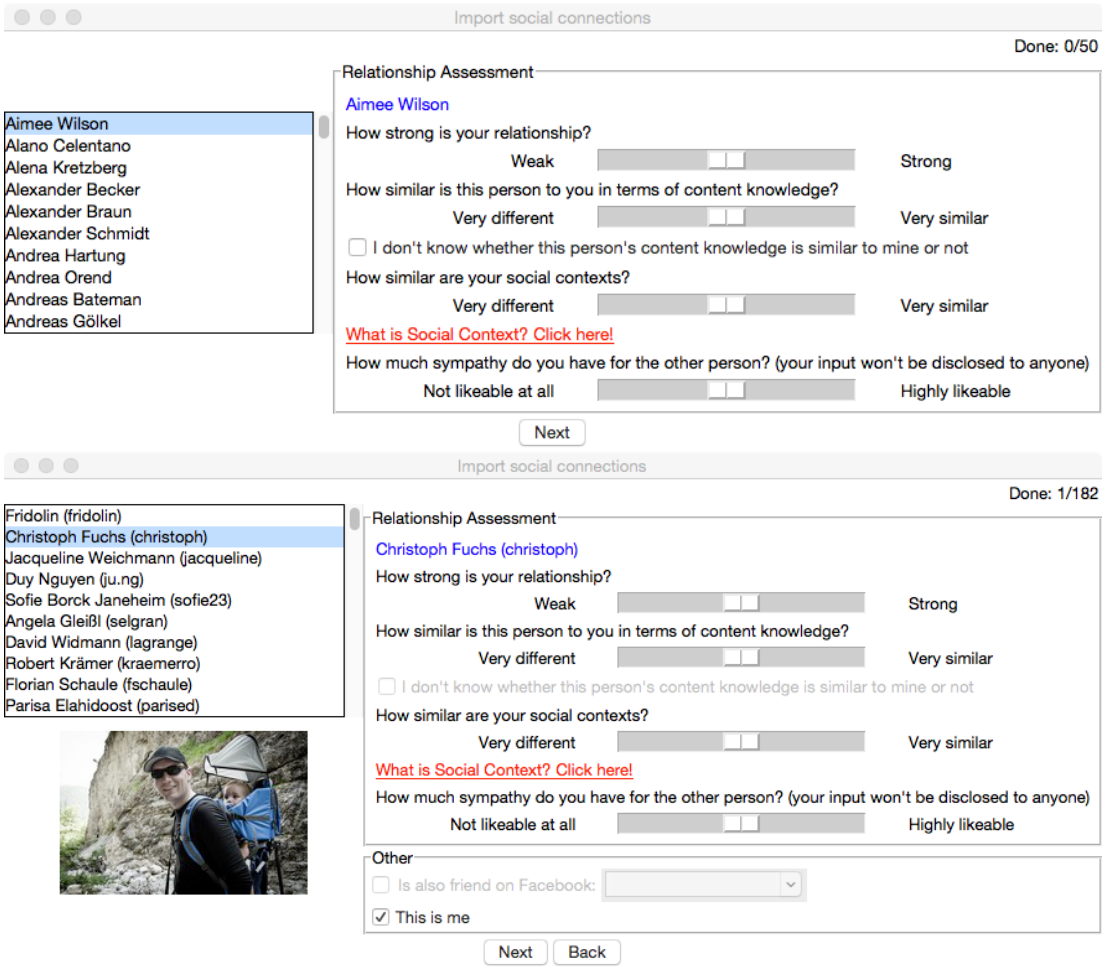


Figure 40: Screenshot of SNExtractor tool (Experiment 7)

based on the standard defined by Pajek (Mrvar, 2015), but stores additional fields for each edge. While the *plain version* contains the names of the nodes in clear text, the names are hashed in the *anonymized version* to protect the participant's privacy. Only the anonymized version is uploaded using the web interface while the plain version needs to be stored on the participant's computer (to be used by SNTranslator, introduced below).

SNTRANSLATOR During Experiment 7a (cf. Section 16.5.2), the participants route queries to certain people in their social network. To preserve the participants' privacy, only the hashed names of the participants' Facebook contacts are stored on the server. To help the participants with the translation, SNTranslator reads the plain version of the social network generated with SNExtractor and calculates the hashed version of the name. This hash can be used in the web interface to document that a specific query has been routed to the contact that is represented by the respective hash value.

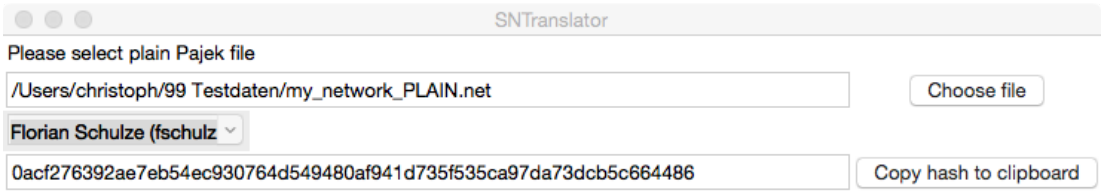


Figure 41: Screenshot of SNTranslator tool (Experiment 7)

GENERATE_WHITELIST, HISTORY_CRAWLER, INDEXER During Experiment 7b (cf. Section 16.5.3), a representation of each participant's information space is required. For several reasons (which are outlined in Section 16.5.3), the information space is based on the content of each participant's visited websites. Therefore, the participants are expected to run the tool generate_whitelist, which extracts the domains of each participant's browsing history and stores them in a plain text file. Each participant has the chance to remove certain domains which should not be considered in the model of the information space. Afterwards, each participant loads her browsing history to her local computer using the history_crawler tool. Once the download finishes, indexer is used to generate topic models with the downloaded websites as input. Each participant uploads the topic models and the dictionaries to the web system.

URLIDENTIFIER To avoid saving each participant's full browsing history on the central server, the URLs of websites considered in each participant's topic model have been mapped to numeric IDs (i.e., instead of uploading the full list of URLs, only the numeric IDs were stored on the central web system). Only if a certain URL has been identified as a relevant response to a query in Experiment 7b (Chapter 16), the participant is asked to provide the URL for the respective numeric ID. The URLIdentifier tool helps to translate the numeric IDs back to actual URLs, relying on a file stored on each participant's computer containing the individual mapping between URLs and numeric IDs.

AMAZONPARSER.JS AmazonParser.js is a JavaScript file that allows the participants to extract all products that have been bought from the Amazon website. It is based on an open source script[1] and has been extended to allow language-specific Amazon versions (e.g. amazon.in or amazon.co.uk). The script is run using the developer console of Firefox or Google Chrome and generates a report. The participants were asked to save the report and analyze it with AmazonBoughtProductsToFile.

AMAZONVIEWEDPRODUCTSTOFILE AmazonViewedProductsToFile is a command-line tool that should be run by the participants to extract the Amazon items that have been viewed, based on each participant's individual browsing history. The tool generates a text file, listing all the identified Amazon items. The participants can manually remove items before uploading the file to the web interface.

1 `https://github.com/CyberLine/amazon-parser` (retrieved 2015-02-03)

AMAZONBOUGHTPRODUCTSTOFILE AmazonBoughtProductsToFile generates a list of items which have been bought from Amazon. The participants can edit the list before uploading it to the web system. The list is based on the output of AmazonParser.js.

BACKEND TOOLS For certain activities taking place in the backend, a small set of tools has been developed:

- *create_pajek_network_from_django* aggregates the various ego-networks uploaded by the participants to one large social network,
- *pajek_anonymizer* applies an additional hashing function to all node names in the newly generated large social network and randomly rewires edges to protect the participant's privacy,
- *detect_in_class_participants* distinguishes active members of the experiment (i.e., actual participants) and social contacts of the participants in the large social network for Experiment 7b (Section 16.5.3, which only includes participants, while 7a also allows information providers who are not a registered participant)
- *neighbor_index_query* identifies a target set of information providers for Experiment 7b (Section 16.5.3) and calculates proposals for the result
- *amazon_alignment* matches the product names extracted from the participant's browsing and purchasing history with Amazon ASIN numbers (using the Amazon API, cf. Chapter 4).

A.2.3 *Additional Results*

VARIABLE	COEFFICIENT	STD. ERROR	T VALUE	P(>|T|)
(Intercept)	43.3700	9.3346	4.65	0.0000
tie strength	0.2632	0.1254	2.10	0.0363
sympathy	-0.3823	0.1367	-2.80	0.0053
social context	-0.5742	0.2591	-2.22	0.0271
$(\text{social context})^2$	0.0109	0.0043	2.55	0.0110
(social context):(content)	0.0056	0.0021	2.63	0.0086
$(\text{social context})^2$:$(\text{tie strength})^2$	-0.0000	0.0000	-2.13	0.0336
$(\text{social context})^2$:$(\text{sympathy})^2$	-0.0000	0.0000	-2.24	0.0255
$(\text{social context})^2$:$(\text{content})^2$	-0.0000	0.0000	-3.05	0.0024
$(\text{social context})^2$:$(\text{tie strength})^2$:$(\text{sympathy})^2$	0.0000	0.0000	2.81	0.0051
$(\text{social context})^2$:$(\text{tie strength})^2$:$(\text{content})^2$	0.0000	0.0000	2.34	0.0199
$(\text{social context})^2$:$(\text{sympathy})^2$:$(\text{content})^2$	0.0000	0.0000	2.97	0.0031
$(\text{social context})^2$:$(\text{tie strength})^2$:$(\text{sympathy})^2$:$(\text{content})^2$	-0.0000	0.0000	-2.98	0.0030

Table 41: Linear model to explain privacy judgment; residuals are not normally distributed and adjusted $R^2 = 0.0413$, F-statistic: 3.171 on 12 and 592 DF, p-value: 0.0002, Residual standard error: 26.38 on 592 degrees of freedom (Experiment 7a)

VARIABLE	COEFFICIENT	STD. ERROR	T VALUE	P(>|T|)
(Intercept)	2.5433	0.3282	7.75	0.0000
social context	-0.0320	0.0159	-2.01	0.0449
$(\text{social context})^2$	0.0007	0.0002	3.34	0.0009
(social context):(content)	0.0004	0.0001	3.40	0.0007
$(\text{social context})^2$:$(\text{content})^2$	-0.0000	0.0000	-4.22	0.0000
$(\text{social context})^2$:$(\text{sympathy})^2$	-0.0000	0.0000	-4.60	0.0000
$(\text{social context})^2$:$(\text{content})^2$:$(\text{sympathy})^2$	0.0000	0.0000	4.02	0.0001
$(\text{social context})^2$:$(\text{content})^2$:$(\text{tie strength})^2$	0.0000	0.0000	2.12	0.0348
$(\text{social context})^2$:$(\text{sympathy})^2$:$(\text{tie strength})^2$	0.0000	0.0000	4.14	0.0000
$(\text{social context})^2$:$(\text{content})^2$:$(\text{sympathy})^2$:$(\text{tie strength})^2$	-0.0000	0.0000	-3.64	0.0003

Table 42: Linear model to explain log(privacy+1); residuals are not normally distributed and adjusted $R^2 = 0.0512$, F-statistic: 4.623 on 9 and 595 DF, p-value: 0.00, Residual standard error: 1.634 on 595 degrees of freedom (Experiment 7a)

VARIABLE	COEFFICIENT	STD. ERROR	Z VALUE	P(>|Z|)
(Intercept)	0.9888	0.3877	2.55	0.0108
tie strength (IS)	0.0094	0.0049	1.90	0.0579
sympathy (IS)	0.0018	0.0055	0.32	0.7466
social context sim. (IS)	-0.0024	0.0040	-0.60	0.5501
content knowledge sim. (IS)	-0.0067	0.0038	-1.75	0.0809

Table 43: Logistic regression model to explain whether information providers reply to requests, Null deviance: 933.07 on 864 degrees of freedom, Residual deviance: 925.64 on 860 degrees of freedom (Experiment 7a)

VARIABLE	COEFFICIENT	STD. ERROR	Z VALUE	P(>|Z|)
(Intercept)	1.0429	0.3044	3.43	0.0006
tie strength (IS)	0.0089	0.0038	2.37	0.0180
content knowledge sim. (IS)	-0.0072	0.0037	-1.94	0.0527

Table 44: Logistic regression model to explain whether information providers reply to requests, Null deviance: 933.07 on 864 degrees of freedom, Residual deviance: 926.05 on 862 degrees of freedom (Experiment 7a)

B

PRIOR PUBLICATIONS

Some ideas, results, and text fragments of this thesis have already appeared previously in the publications listed on page vii. To improve the readability of thesis, direct quotations referencing to these publications have not been marked accordingly in the respective sections. The following table shows in detail which parts have already appeared in which publication linked to the thesis.

Section and text in thesis	Prior publication	Type
Chapter 1 ("Today's prevailing (...) with the community"; "With the rise of (...) generalization of the results"; research questions)	(Fuchs and Groh, 2016a) (Section I)	direct quotation (adjusted references)
Section 2.5.1 ("Borgatti and Cross (...) too costly")	(Fuchs and Groh, 2016a) (Section II)	direct quotation (adjusted references)
Exp. 1, Section 10.1	(Fuchs et al., 2015) (Abstract)	direct quotation
Exp. 1, Section 10.2	(Fuchs et al., 2015) (Introduction)	direct quotation, with minimal adjustments (e.g., cited literature and references)
Exp. 1, Section 10.3	(Fuchs et al., 2015) (Section 2.1)	direct quotation
Exp. 1, Section 10.4	(Fuchs et al., 2015) (Section 2.2)	direct quotation
Exp. 1, Section 10.5	(Fuchs et al., 2015) (Section 2.3)	direct quotation
Exp. 1, Section 10.6	(Fuchs et al., 2015) (Section 3)	direct quotation
Exp. 1, Section 10.7	(Fuchs et al., 2015) (Section 4)	direct quotation
Exp. 1, Figure 5	(Fuchs et al., 2015) (Figure 1a)	slightly adjusted presentation
Exp. 1, Table 5	(Fuchs et al., 2015) (Figure 1b)	slightly adjusted presentation
Exp. 2, Section 11.1	(Fuchs and Groh, 2015b) (Abstract)	direct quotation, with minimal adjustments (added reference, appended "for the information seeker")
Exp. 2, Section 11.2 (first paragraph: "Previous results (...) response quality.")	(Fuchs and Groh, 2015b) (Section I, paragraph 5)	direct quotation
Exp. 2, Section 11.2 (second paragraph: "A laboratory experiment (...) than normal status messages.")	(Fuchs and Groh, 2015b) (Section IIb, paragraph 3)	direct quotation
Exp. 2, Section 11.3	(Fuchs and Groh, 2015b) (Section IIIa)	direct quotation, added sentence "A positive result would suggest that the routing of questions in a social (...) potential information providers?)."
Exp. 2, Section 11.4	(Fuchs and Groh, 2015b) (Section IIIb)	direct quotation
Exp. 2, Section 11.5	(Fuchs and Groh, 2015b) (Section IIIc)	direct quotation
Exp. 2, Table 6	(Fuchs and Groh, 2015b) (Table 1)	direct quotation
Exp. 2, Section 11.6	(Fuchs and Groh, 2015b) (Section IV)	direct quotation, omitted last paragraph (Section IVd)
Exp. 2, Section 11.7	(Fuchs and Groh, 2015b) (Section V)	direct quotation

Exp. 2, Figure 7	(Fuchs and Groh, 2015b) (Figure 2)	direct quotation
Exp. 3, Section 12.1	(Fuchs et al., 2016a) (Abstract)	direct quotation
Exp. 3, Section 12.2	(Fuchs et al., 2016a) (Section I)	direct quotation
Exp. 3, Section 12.3 (first paragraph)	(Fuchs et al., 2016a) (Section II)	direct quotation
Exp. 3, Section 12.4	(Fuchs et al., 2016a) (Section IV)	direct quotation (adjusted references)
Exp. 3, Section 12.5	(Fuchs et al., 2016a) (Section V)	direct quotation (adjusted references)
Exp. 3, Section 12.6	(Fuchs et al., 2016a) (Section VI, Section VIII)	direct quotation (adjusted references)
Exp. 3, Section 12.7	(Fuchs et al., 2016a) (Section VII)	direct quotation (adjusted references)
Exp. 3, Figure 8	(Fuchs et al., 2016a) (Figure 1)	direct quotation
Exp. 3, Table 7	(Fuchs et al., 2016a) (Table V)	direct quotation
Exp. 3, Table 8	(Fuchs et al., 2016a) (Table VI)	direct quotation
Exp. 3, Table 9	(Fuchs et al., 2016a) (Table I)	direct quotation
Exp. 3, Table 10	(Fuchs et al., 2016a) (Table II)	direct quotation
Exp. 4, Section 13.1	(Fuchs and Groh, 2016b) (Abstract)	direct quotation (adjusted references)
Exp. 4, Section 13.2	(Fuchs and Groh, 2016b) (Section I)	direct quotation (adjusted references)
Exp. 4, Section 13.3	(Fuchs and Groh, 2016b) (Section I)	direct quotation (adjusted references)
Exp. 4, Section 13.4	(Fuchs and Groh, 2016b) (Section III)	direct quotation (adjusted references)
Exp. 4, Section 13.5	(Fuchs and Groh, 2016b) (Section IV)	direct quotation (adjusted references)
Exp. 4, Section 13.6	(Fuchs and Groh, 2016b) (Section V)	direct quotation (adjusted references)
Exp. 4, Section 13.7	(Fuchs and Groh, 2016b) (Section VI)	direct quotation (adjusted references, corrected: 10 topics for LDA)
Exp. 4, Figure 9	(Fuchs and Groh, 2016b) (Figure 1)	direct quotation (adjusted references)
Exp. 4, Figure 10	(Fuchs and Groh, 2016b) (Figure 2)	direct quotation
Exp. 5, Section 14.1	(Fuchs et al., 2016b) (Abstract)	direct quotation (adjusted references)
Exp. 5, Section 14.2	(Fuchs et al., 2016b) (Section I)	direct quotation, with small adjustments (adjusted references, added information why LDA has been choosen)
Exp. 5, Section 14.3	(Fuchs et al., 2016b) (Section I)	direct quotation
Exp. 5, Section 14.4	(Fuchs et al., 2016b) (Section III)	direct quotation (adjusted references)

Exp. 5, Section 14.5	(Fuchs et al., 2016b) (Section IV)	direct quotation (adjusted references)
Exp. 5, Section 14.6	(Fuchs et al., 2016b) (Section V, VII)	direct quotation (adjusted references)
Exp. 5, Section 14.8	(Fuchs et al., 2016b) (Section VI)	direct quotation (adjusted references)
Exp. 5, Table 11	(Fuchs et al., 2016b) (Table I)	direct quotation
Exp. 5, Figure 11	(Fuchs et al., 2016b) (Figure 1)	direct quotation
Exp. 5, Figure 12	(Fuchs et al., 2016b) (Figure 2)	direct quotation
Exp. 5, Figure 13	(Fuchs et al., 2016b) (Figure 3)	direct quotation
Exp. 5, Figure 14	(Fuchs et al., 2016b) (Figure 5)	direct quotation
Exp. 5, Figure 15	(Fuchs et al., 2016b) (Figure 6)	direct quotation
Exp. 5, Figure 16	(Fuchs et al., 2016b) (Figure 7)	direct quotation
Exp. 6, Section 15.2 (“The underlying hypothesis (...) similar to IDF.”)	(Fuchs et al., 2016b) (Section IVc)	direct quotation
Exp. 6, Section 15.5 (“To calculate the degree of (...) using cosine similarity.”)	(Fuchs et al., 2016b) (Section IVc)	direct quotation
Exp. 6, Section 15.6 (“ESA performs better (...) content perspective).”)	(Fuchs et al., 2016b) (Section Vb)	direct quotation
Exp. 6, Section 15.7 (“Especially in the (...) demonstrate their strengths.”)	(Fuchs et al., 2016b) (Section VI)	direct quotation
Exp. 6, Figure 17	(Fuchs et al., 2016b) (Figure 4)	direct quotation
Exp. 7, Section 16.5.1	(Fuchs and Groh, 2016a) (Section IIIb)	indirect quotation, enriched with additional information, and adjusted references
Exp. 7, Table 12	(Fuchs and Groh, 2016a) (Table I)	direct quotation
Exp. 7, Section 16.5.2	(Fuchs and Groh, 2016a) (Section IIIc)	direct quotation (adjusted references)
Exp. 7, Section 16.5.4 (“The participants (...) rating interface”)	(Fuchs and Groh, 2016a) (Section IIId)	direct quotation (adjusted references)
Exp. 7, Table 14	(Fuchs and Groh, 2016a) (Table II)	direct quotation
Exp. 7, Section 16.6.1.1	(Fuchs and Groh, 2016a) (Section IVa)	direct quotations, thesis provides more details, adjusted references
Exp. 7, Figure 18	(Fuchs and Groh, 2016a) (Fig. 1)	direct quotation
Exp. 7, Table 43	(Fuchs and Groh, 2016a) (Table III)	direct quotation

Exp. 7, Table 44	(Fuchs and Groh, 2016a) (Table IV)	direct quotation
Exp. 7, Table 15	(Fuchs and Groh, 2016a) (Table V)	direct quotation
Exp. 7, Section 16.6.1.2	(Fuchs and Groh, 2016a) (Section IVb)	direct quotations, thesis provides more details, adjusted references
Exp. 7, Table 17	(Fuchs and Groh, 2016a) (Table VI)	direct quotation
Exp. 7, Table 20	(Fuchs and Groh, 2016a) (Table VII)	direct quotation
Exp. 7, Table 21	(Fuchs and Groh, 2016a) (Table VIII)	direct quotation
Exp. 7, Figure 21	(Fuchs and Groh, 2016a) (Fig. 2)	direct quotation
Exp. 7, Figure 22	(Fuchs and Groh, 2016a) (Fig. 3)	direct quotation
Exp. 7, Figure 23	(Fuchs and Groh, 2016a) (Fig. 4)	direct quotation
Exp. 7, Table 22	(Fuchs and Groh, 2016a) (Table IX)	direct quotation
Exp. 7, Table 23	(Fuchs and Groh, 2016a) (Table X)	direct quotation
Exp. 7, Table 24	(Fuchs and Groh, 2016a) (Table XI)	direct quotation
Exp. 7, Section 16.6.1.3	(Fuchs and Groh, 2016a) (Section IVc)	direct quotations, thesis provides more details
Exp. 7, Table 25	(Fuchs and Groh, 2016a) (Table XII)	direct quotation
Exp. 7, Table 26	(Fuchs and Groh, 2016a) (Table XIII)	direct quotation
Exp. 7, Figure 25	(Fuchs and Groh, 2016a) (Fig. 5)	direct quotation
Exp. 7, Table 30	(Fuchs and Groh, 2016a) (Table XIV)	direct quotation
Exp. 7, Section 16.6.1.5	(Fuchs and Groh, 2016a) (Section IVd)	direct quotations, thesis provides more details, adjusted references
Exp. 7, Table 31	(Fuchs and Groh, 2016a) (Table XV)	direct quotation
Exp. 7, Figure 28	(Fuchs and Groh, 2016a) (Fig. 6)	direct quotation
Exp. 7, Table 32	(Fuchs and Groh, 2016a) (Table XVI)	direct quotation
Exp. 7, Section 16.6.1.6	(Fuchs and Groh, 2016a) (Section IVe)	direct quotations, thesis provides more details, adjusted references
Exp. 7, Figure 33	(Fuchs and Groh, 2016a) (Fig. 7)	direct quotation
Exp. 7, Figure 30	(Fuchs and Groh, 2016a) (Fig. 8)	direct quotation
Exp. 7, Table 33	(Fuchs and Groh, 2016a) (Table XVII)	direct quotation

Exp. 7, Section 16.6.2.1	(Fuchs and Groh, 2016a) (Section Va)	direct quotations, thesis provides more details, adjusted references
Exp. 7, Figure 32	(Fuchs and Groh, 2016a) (Fig. 9)	direct quotation
Exp. 7, Figure 35	(Fuchs and Groh, 2016a) (Fig. 10)	direct quotation
Exp. 7, Figure 36	(Fuchs and Groh, 2016a) (Fig. 11)	direct quotation
Exp. 7, Section 16.7	(Fuchs and Groh, 2016a) (Section VI)	direct quotations, thesis provides more details, adjusted references
Exp. 7, Figure 37	(Fuchs and Groh, 2016a) (Fig. 12)	direct quotation
Exp. 7, Figure 38	(Fuchs and Groh, 2016a) (Fig. 13)	direct quotation
Exp. 7, Chapter 18 (introduction and paragraphs 1, 2, 6, 7)	(Fuchs and Groh, 2016a) (Section VII)	direct quotation
Exp. 7, Table 38	(Fuchs and Groh, 2016a) (Table XVIII)	direct quotation
Exp. 7, Table 13	(Fuchs and Groh, 2016a) (Table XIX)	direct quotation

Table 45: Mapping table for quotations from prior publications

BIBLIOGRAPHY

Adler, P. S. and Kwon, S.-W. (2002). Social Capital: Prospects for a New Concept. *The Academy of Management Review*, 27(1):17–40.

Agichtein, E., Brill, E., and Dumais, S. (2006). Improving Web Search Ranking by Incorporating User Behavior Information. In *Proceedings of the 29th Annual International Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 19–26, New York, NY, USA. ACM.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding High-quality Content in Social Media. In *Proceedings of the International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194, New York, NY, USA. ACM.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc.

Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. (2011). Social fMRI: Investigating and Shaping Social Mechanisms in the Real World. *Pervasive and Mobile Computing*, 7:643–659.

Alavi, M. and Leidner, D. E. (2001). Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, 25(1):107–136.

Arguello, J. (2011). *Federated Search for Heterogeneous Environments*. PhD thesis, Carnegie Mellon University.

Azzopardi, L. (2011). The Economics in Interactive Information Retrieval. In *Proceedings of the 34th International Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 15–24, New York, NY, USA. ACM.

Azzopardi, L. (2014). Modelling Interaction with Economic Models of Search. In *Proceedings of the 37th International Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 3–12, New York, NY, USA. ACM.

Azzopardi, L., Kelly, D., and Brennan, K. (2013). How Query Cost Affects Search Behavior. In *Proceedings of the 36th International Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 23–32, New York, NY, USA. ACM.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*. ACM Press and Addison Wesley.

Baillie, M., Carman, M., and Crestani, F. (2011). A Multi-collection Latent Topic Model for Federated Search. *Information Retrieval*, 14(4):390–412.

Balog, K. and de Rijke, M. (2007). Determining Expert Profiles (With an Application to Expert Finding). In *Proceedings of International Joint Conference on Artificial Intelligence*, IJCAI '07, pages 2657–2662.

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., and Si, L. (2012). Expertise Retrieval. *Foundations and Trends in Information Retrieval*, 6(2-3):127–256.

Becerra-Fernandez, I. (2006). Searching for Experts on the Web: A Review of Contemporary Expertise Locator Systems. *ACM Transactions on Internet Technologies*, 6(4):333–355.

Berk, J. and DeMarzo, P. (2014). *Corporate Finance*. Pearson, third edition.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43.

Bertino, E. and Matei, S. A., editors (2015). *Roles, Trust, and Reputation in Social Media Knowledge Markets*. Springer International Publishing.

Beyer, H. and Holtzblatt, K. (1998). *Contextual Design. Defining Customer-Centered Systems*. Academic Press.

Bhattacharyya, P. and Wu, S. F. (2014). InfoSearch: A Social Search Engine. In Chu, W. W., editor, *Data Mining and Knowledge Discovery for Big Data*, pages 193–223. Springer Berlin Heidelberg.

Birnkammerer, S. (2010). Design of a Protocol for Flow-Control of Information-Items in Decentralized Social Networking. Master's thesis, TU München. Supervised by Georg Groh.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M. and Lafferty, J. D. (2009). Topic Models. In Srivastava, A. and Sahami, M., editors, *Text Mining: Classification, Clustering, and Applications*, pages 71–93. Taylor & Franics Group.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Boettcher, C. (2016). Sharing Knowledge in Social Information Retrieval Scenarios. Master's thesis, Technische Universität München. Supervised by Christoph Fuchs and Georg Groh.

Bordino, I., Mejova, Y., and Lalmas, M. (2013). Penguins in Sweaters, or Serendipitous Entity Search on User-Generated Content. In *Proceedings of the 22nd International Conference on Information & Knowledge Management*, CIKM '13, pages 109–118, New York, NY, USA. ACM.

Borgatti, S. P. and Cross, R. (2003). A Relational View of Information Seeking and Learning in Social Networks. *Management Science*, 49(4):432–445.

Breusch, T. S. and Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294.

Brin, S. and Page, L. (1999). The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30:107–117.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference - Understanding AIC and BIC in Model Selection. *Sociological Methods and Research*, 33:261–304.

Burnkrant, R. E. and Cousineau, A. (1975). Normative Social Influence in Buyer Behavior. *Journal of Consumer Research*, 2(3):206–215.

Callan, J. (2000). Distributed Information Retrieval. *The Information Retrieval Series*, 7:127–150.

Cao, X., Cong, G., and Jensen, C. S. (2010). Mining Significant Semantic Locations From GPS Data. In *Proceedings of the VLDB Endowment*, volume 3, pages 1009–1020.

Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'el, N., Ronen, I., Uziel, E., Yogev, S., and Chernov, S. (2009). Personalized Social Search Based on the User's Social Network. In *Proceedings of the 18th Conference on Information and Knowledge Management*, CIKM '09, pages 1227–1236, New York, NY, USA. ACM.

Chaum, D. L. (1981). Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. *Communications of the ACM*, 24(2):84–90.

Chen, G. and Kotz, D. (2005). A Survey of Context-Aware Mobile Computing Research. *Dartmouth Computer Science Technical Report*, TR2000-381.

Chen, H., Finin, T., and Joshi, A. (2005). The SOUPA Ontology for Pervasive Computing. In *Ontologies for Agents: Theory and Experience Whitestein Series in Software Agent Technologies*, pages 233–258.

Chen, J. R., Wolfe, S. R., and Wragg, S. D. (2000). A Distributed Multi-Agent System for Collaborative Information Management and Sharing. In *Proceedings of the 9th International Conference on Information and Knowledge Management*, CIKM '00, pages 382–388. ACM.

Chi, E. H., Pirolli, P., Chen, K., and Pitkow, J. (2001). Using Information Scent to Model User Information Needs and Actions and the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 490–497, New York, NY, USA. ACM.

Christen, M. (2015). YaCy Decentralized Web Search. `http://yacy.net/en/index.html` (retrieved 2015-07-21).

Christensen, R. H. B. (2015a). A Tutorial on Fitting Cumulative Link Models with the Ordinal Package. `https://cran.r-project.org/web/packages/ordinal/vignettes/clm_tutorial.pdf` (retrieved 2016-01-16).

Christensen, R. H. B. (2015b). Regression Models for Ordinal Data. https://cran.r-project.org/web/packages/ordinal/ordinal.pdf (retrieved 2015-11-24).

Church, K., Cousin, A., and Oliver, N. (2012). I wanted to settle a bet!: understanding why and how people use mobile search in social settings. *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*.

Church, K. and Oliver, N. (2011). Understanding Mobile Web and Mobile Search Use in Today's Dynamic Mobile Landscape. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 67–76, New York, NY, USA. ACM.

Church, K. and Smyth, B. (2009). Understanding the Intent Behind Mobile Information Needs. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, pages 247–256, New York, NY, USA. ACM.

Church, K., Smyth, B., Cotter, P., and Bradley, K. (2007). Mobile Information Access: A Study of Emerging Search Behavior on the Mobile Internet. *ACM Transactions on the Web*, 1(1).

Cimiano, P., Schultz, A., Sizov, S., Sorg, P., and Staab, S. (2009). Explicit Versus Latent Concept Models for Cross-language Information Retrieval. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI '09, pages 1513–1518, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46.

Constant, D., Sproull, L., and Kiesler, S. (1996). The Kindness of Strangers: The Usefulness of Electronic Weak Ties for Technical Advice. *Organization Science*, 7(2):119–135.

Crawley, M. J. (2007). *The R Book*. Wiley.

de Vries, A. (2015). Recommendation and Information Retrieval: Two Sides of the Same Coin? 10th European Summer School in Information Retrieval, http://mklab.iti.gr/essir2015/wp-content/uploads/2015/03/ESSIR2015_Vries.pdf (retrieved 2016-03-01).

Dearman, D., Kellar, M., and Truong, K. N. (2008). An Examination of Daily Information Needs and Sharing Opportunities. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dey, A. K. (2001). Understanding and Using Context. *Personal Ubiquitous Comput.*, 5(1):4–7.

Dong, H., Hussain, F. K., and Chang, E. (2008). A Survey in Semantic Search Technologies. In *Proceedings of the Second International Conference on Digital Ecosystems and Technology*, pages 403–408, Phitsanulok, Thailand. IEEE.

Dörk, M., Carpendale, S., and Williamson, C. (2011). The Information Flaneur: A Fresh Look at Information Seeking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1215–1224, New York, NY, USA. ACM.

Dörk, M., Riche, N. H., Ramos, G., and Dumais, S. (2012). PivotPaths: Strolling Through Faceted Information Spaces. In *IEEE Transactions On Visualization and Computer Graphics*, volume 18 of *InfoVis '12*, pages 2709–2718. IEEE.

Duggan, G. B. and Payne, S. J. (2008). Knowledge in the Head and on the Web: Using Topic Expertise to Aid Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 39–48.

Dunbar, R. I. M. (1992). Neocortex Size as a Constraint on Group Size in Primates. *Journal of Human Evolution*, 22(6):469–493.

Dunbar, R. I. M. (1993). Coevolution of Neocortical Size, Group Size and Language in Humans. *Behavioral and Brain Sciences*, 16(4):681– 694.

Durbin, J. and Watson, G. S. (1971). Testing for Serial Correlation in Least Squares Regression. III. *Biometrika*, 58(1):1–19.

Eagle, N. and Pentland, A. (2006). Reality Mining: Sensing Complex Social Systems. *Journal of Personal and Ubiquitous Computing*, 10(4).

Efron, M. and Golovchinsky, G. (2011). Estimation Methods for Ranking Recent Information. In *Proceedings of the 34th International Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 495–504, New York, NY, USA. ACM.

Eickhoff, C., Teevan, J., White, R., and Dumais, S. (2014). Lessons from the Journey: A Query Log Analysis of Within-Session Learning. In *Proceedings of the 7th International Conference on Web Search and Data Mining*. ACM.

ElGamal, T. (1985). A Public-Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms. *IEEE Transactions on Information Theory*, 31(4):469–472.

Endres, D. M. and Schindelin, J. E. (2003). A New Metric for Probability Distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.

Evans, B. M., Kairam, S., and Pirolli, P. (2009). Exploring the Cognitive Consequences of Social Search. In *CHI Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, pages 3377–3382, New York, NY, USA. ACM.

Fiske, A. P. (1992). The Four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations. *Psychological Review*, 99(4):689–723.

Foner, L. N. (1997). Yenta: A Multi-agent, Referral-based Matchmaking System. In *Proceedings of the First International Conference on Autonomous Agents*, AGENTS '97, pages 301–307, New York, NY, USA. ACM.

Forte, A., Dickard, M., Magee, R., and Agosto, D. E. (2014). What Do Teens Ask Their Online Social Networks? Social Search Practices among High School Students. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 28–37. ACM.

Franchi, E., Poggi, A., and Tomaiuolo, M. (2013). Supporting Social Networks With Agent-Based Services. *International Journal of Virtual Communities and Social Networking*, 5(1):62–74.

Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys*, 42(4):14:1–14:53.

Gabrilovich, E. and Markovitch, S. (2009). Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research*, 34:443–498.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Ghorab, M. R., Zhou, D., O'Connor, A., and Wade, V. (2013). Personalised Information Retrieval: Survey and Classification. *User Modeling and User-Adapted Interaction*, 23(4):381–443.

Gilbert, E. and Karahalios, K. (2009). Predicting Tie Strength With Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 211–220.

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2009). A Walk in Facebook: Uniform Sampling of Users in Online Social Networks. *arXiv:0906.0060 [cs.SI]*. `http://arxiv.org/pdf/0906.0060.pdf` (retrieved 2016-01-17).

Göker, A., editor (2009). *Information Retrieval: Searching in the 21st Century*. Wiley, 1. edition.

Goldberg, L. R. (1993). The Structure of Phenotypic Personality Traits. *American Psychologist*, 48:26–34.

Golder, S. A. and Yardi, S. (2010). Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality. In *Proceedings of the 2nd International Conference on Social Computing*, SocialCom '10, pages 88–95. IEEE.

Gou, L., Zhang, X. L., Chen, H.-H., Kim, J.-H., and Giles, C. L. (2010). Social Network Document Ranking. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 313–322, New York, NY, USA. ACM.

Govaerts, S., El Helou, S., Duval, E., and Gillet, D. (2011). A Federated Search and Social Recommendation Widget. In *Proceedings of the 2nd International Workshop on Social Recommender Systems (SRS 2011) in conjunction with the 2011 ACM Conference on Computer Supported Cooperative Work (CSCW 2011)*.

Granovetter, M. (1976). Network Sampling: Some First Steps. *The American Journal of Sociology*, 81(6):1287–1303.

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, 1:201–233.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380.

Groh, G. (2011). *Contextual Social Networking*. Habilitation, TU München.

Groh, G. and Birnkammerer, S. (2011). Privacy and Information Markets: Controlling Information Flows in Decentralized Social Networking. In *Proceedings of the 3rd International Conference on Social Computing*, SocialCom '11, pages 856–861. IEEE.

Groh, G., Brocco, M., and Kleemann, A. (2011a). Interest-Based vs. Social Person-Recommenders in Social Networking Platforms. *arXiv:1107.5654 [cs.SI]*. `http://arxiv.org/pdf/1107.5654v1` (retrieved 2016-01-17).

Groh, G. and Ehmig, C. (2007). Recommendations in Taste Related Domains: Collaborative Filtering vs. Social Filtering. In *Proceedings of the International Conference on Supporting Group Work*, GROUP '07. ACM.

Groh, G., Fuchs, C., and Lehmann, A. (2011b). Combining Evidence for Social Situation Detection. In *Proceedings of the 3rd International Conference on Social Computing*, SocialCom '11, pages 742–747. IEEE.

Groh, G., Lehmann, A., and de Souza, M. (2011c). Mobile Detection of Social Situations with Turn-Taking Patterns. In *Proceedings of International Conferences Informatics*, pages 137–142. IADIS.

Groh, G., Lehmann, A., Reimers, J., Frieß, M. R., and Schwarz, L. (2010). Detecting Social Situations from Interaction Geometry. In *Proceedings of the Second International Conference on Social Computing*, SocialCom '10.

Groh, G., Straub, F., Eicher, J., and Grob, D. (2013). Geographic Aspects of Tie Strength and Value Of Information in Social Networking. In *Proceedings of the 6th SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 1–10. ACM.

Groza, T., Handschuh, S., Möller, K., Grimnes, G., Sauermann, L., Minack, E., Mesnage, C., Jazayeri, M., Reif, G., and Gudjónsdottir, R. (2007). The NEPOMUK Project - On the way to the Social Semantic Desktop. In *Proceedings of I-SEMANTICS*, Graz, Austria.

Gulli, A., Cataudella, S., and Foschini, L. (2009). TC-SocialRank: Ranking the Social Web. In Avrachenkov, K., Donato, D., and Litvak, N., editors, *Algorithms and Models for the Web-Graph: 6th International Workshop*, WAW, pages 143–154. Springer Berlin Heidelberg.

Han, J., Schmidtke, H. R., Xie, X., and Woo, W. (2013). Adaptive Content Recommendation for Mobile Users: Ordering Recommendations Using a Hierarchical Context Model With Granularity. *Pervasive and Mobile Computing*.

Hangal, S. (2012). *Reshaping Reminiscence, Web Browsing and Web Search Using Personal Digital Archives*. PhD thesis, Stanford University.

Hangal, S., Lam, M. S., and Heer, J. (2011). MUSE: Reviving Memories Using Email Archives. In *Proceedings of the 24th Symposium on User Interface Software and Technology*, UIST '11, pages 75–84. ACM.

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. (2013). Measuring Personalization of Web Search. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 527–538.

Hansen, M. T. (1999). The Search-transfer Problem: The Role of Weak Ties in Sharing Knowledge Across Organization Subunits. *Administrative Science Quarterly*, 44(1):82–111.

Hartl, F. (2013). Topic Recommender Systems in Social Networks Using Topic Models. Master's thesis, Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Haveliwala, T. H. (2002). Topic-Sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 517–526, New York, NY, USA. ACM.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1):75–105.

Heymann, P., Koutrika, G., and Garcia-Molina, H. (2008). Can Social Bookmarking Improve Web Search? In *Proceedings of the International Conference on Web Search and Data Mining*, WSDM '08, pages 195–206, New York, NY, USA. ACM.

Hiemstra, D. (2009). *Information Retrieval Models*. Wiley.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.

Horowitz, D. and Kamvar, S. D. (2010). The Anatomy of a Large-scale Social Search Engine. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 431–440.

Hsiao, I.-H., Bakalov, F., Brusilovsky, P., and König-Ries, B. (2013). Progressor: Social Navigation Support Through Open Social Student Modeling. *New Review of Hypermedia and Multimedia*, 19(2).

Huang, H., Gao, Y., Chen, L., Li, R., Chiew, K., and He, Q. (2013). Browse with a social web directory. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 865–868, New York, NY, USA. ACM.

i Mansilla, A. T. and de la Rosa i Esteva, J. L. (2013). Survey of Social Search from the Perspectives of the Village Paradigm and Online Social Networks. *Journal of Information Science*, 39(5):688–707.

Ipeirotis, P. G. and Gravano, L. (2002). Distributed Search Over the Hidden Web: Hierarchical Database Sampling and Selection. In *Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB '02, pages 394–405.

Joung, Y.-J., Chen, S.-M., Wu, C.-C., and Chiu, T.-Y. (2013). A Comparative Study of Expert Search Strategies in Online Social Networks. In *Proceedings of the 27th International Conference on Advanced Information Networking and Applications*, AINA '13, pages 960–967.

Kahanda, I. and Neville, J. (2009). Using Transactional Information to Predict Link Strength in Online Social Networks. In *Proceedings of the 3rd International Conference on Weblogs and Social Media*, ICWSM '09, pages 74–81. AAAI.

Kamvar, M. and Baluja, S. (2006). A Large Scale Study of Wireless Search Behavior: Google Mobile Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 701–709. ACM.

Kamvar, M., Kellar, M., Patel, R., and Xu, Y. (2009). Computers and iPhones and Mobile Phones, Oh My!: A Logs-based Comparison of Search Users on Different Devices. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 801–810.

Kautz, H., Selman, B., and Shah, M. (1997a). Referral Web: Combining Social Networks and Collaborative Filtering. *Communications of the ACM*, 40(3):63–65.

Kautz, H., Selman, B., and Shah, M. (1997b). The Hidden Web. *AI Magazine*, 18(2):27.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. (2011). Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media. *Business Horizons*, 54(3):241 – 251.

Kissner, L. and Song, D. (2005). Privacy-preserving Set Operations. *Lecture Notes in Computer Science*, 3621:241–257.

Kleinberg, J. (2000). Navigation in a Small World. *Nature*, 406:845.

Kleinberg, J. (2006a). Complex Networks and Decentralized Search Algorithms. In *Proceedings of the International Congress of Mathematicians*.

Kleinberg, J. (2006b). Social Networks, Incentives, and Search. In *Proceedings of the 29th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 210–211. ACM.

Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632.

Klimt, B. and Yang, Y. (2004). Introducing the Enron Corpus. In *Proceedings of the Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '04.

Kontominas, D., Raftopoulou, P., Tryfonopoulos, C., and Petrakis, E. G. (2013). DS4: A Distributed Social and Semantic Search System. In *Advances in Information Retrieval – 35th European Conference on IR Research*, ECIR '13, pages 832–836. Springer.

Kossinets, G., Kleinberg, J., and Watts, D. (2008). The Structure of Information Pathways in a Social Communication Network. In *Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 435–443, New York, NY, USA. ACM.

Koster, B. (2013). Modeling Influence in Social Networks with Topic Models. Master's thesis, Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Kramár, T. and Bilevic, R. (2014). Modeling the Dynamics of Research Interests. *Cognitive Traveling in Digital Space of the Web and Digital Libraries: Yield of the Interdisciplinary Multi-partner Project TraDice*. `http://tradice.fiit.stuba.sk/proc/tradice2013_submission_13.pdf` (retrieved 2016-01-02).

Kruskal, W. H. and Wallis, W. A. (1952). Use of Ranks in One-criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Kukla, G., Kazienko, P., Bródka, P., and Filipowski, T. (2012). SocLaKE: Social Latent Knowledge Explorator. *The Computer Journal*, 55(3).

Lampe, C., Gray, R., Fiore, A. T., and Ellison, N. (2014). Help is on the Way: Patterns of Responses to Resource Requests on Facebook. In *Proceedings of the 17th Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 3–15. ACM.

Lappas, T., Liu, K., and Terzi, E. (2011). A Survey of Algorithms and Systems for Expert Location in Social Networks. In Aggarwal, C. C., editor, *Social Network Data Analytics*, pages 215–241. Springer.

Lee, D. H. and Brusilovsky, P. (2012). Exploring Social Approach to Recommend Talks at Research Conferences. In *Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, CollaborateCom '12, pages 157–164. IEEE.

Levin, D. Z. and Cross, R. (2004). The Strength of Weak Ties You Can Trust: The Mediating Role of Trust in Effective Knowledge Transfer. *Management Science*, 50(11):1477–1490.

Li, X. and Chang, S.-K. (2007). User Profiling in the Chronobot/Virtual Classroom System. *International Journal of Software Engineering and Knowledge Engineering*, 17(2):191–206.

Li, X. and Croft, W. B. (2003). Time-based Language Models. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, CIKM '03, pages 469–475, New York, NY, USA. ACM.

Lipford, H. R., Besmer, A., and Watson, J. (2008). Understanding Privacy Settings in Facebook with an Audience View. In *Proceedings of the 1st Conference on Usability, Psychology, and Security*, UPSEC'08, pages 2:1–2:8, Berkeley, CA, USA. USENIX Association.

Liu, J., Wolfson, O., and Yin, H. (2006). Extracting Semantic Location from Outdoor Positioning Systems. In *Proceedings of the 7th International Conference on Mobile Data Management*, MDM '06, page 73.

Llanes, O. B. (2016). Comparing Means of Information Retrieval Based on Explicit Concepts and Latent Topics for Social Search. Master's thesis, Technische Universität München. Supervised by Christoph Fuchs and Georg Groh.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables (Advanced Quantitative Techniques in the Social Sciences)*. Sage Publications.

Lu, D. and Li, Q. (2011). Personalized Search on Flickr Based on Searcher's Preference Prediction. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 81–82, New York, NY, USA. ACM.

Lu, J. (2007). *Full-Text Federated Search in Peer-to-Peer Networks*. PhD thesis, Carnegie Mellon University.

Macdonald, C. and Ounis, I. (2006). Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task. In *Proceedings of the 15th International Conference on Information and Knowledge Management*, CIKM '06, pages 387–396, New York, NY, USA. ACM.

Madan, A., Farrahi, K., Gatica-Perez, D., and Pentland, A. (2011). Pervasive Sensing to Model Political Opinion in Face-to-Face Networks. *Lecture Notes in Computer Science*, 6696:214–231.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mari, M., Poggi, A., Tomaiuolo, M., and Turci, P. (2006). Enhancing Information Sharing Through Agents. In *Proceedings of the 8th International BI Conference on Agent-oriented Information Systems IV*, pages 202–211.

Maron, M. E. and Kuhns, J. L. (1960). On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM*, 7(3):216–244.

McCay-Peet, L. and Toms, E. (2011). Measuring the Dimensions of Serendipity in Digital Environments. *Information Research: An International Electronic Journal*, 16(3).

McDonnell, M. and Shiri, A. (2011). Social Search: A Taxonomy of, and a User-Centred Approach to, Social Web Search. *Program: Electronic Library and Information Systems*, 45(1):6–28.

Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. (2007). Personalized Search on the World Wide Web. *Lecture Notes in Computer Science*, 4321:195–230.

Micarelli, A. and Sciarrone, F. (2004). Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System. *User Modeling and User-Adapted Interaction*, 14(2):159–200.

Mihajlovic, V. (2006). *Score Region Algebra – A Flexible Framework for Structured Information Retrieval*. PhD thesis, University of Twente.

Milgram, S. (1967). The Small World Problem. *Psychology Today*, 1(1):61–67.

Millen, D. R., Yang, M., Whittaker, S., and Feinberg, J. (2007). Social Bookmarking and Exploratory Search. In *Proceedings of the 10th European Conference on Computer-Supported Cooperative Work*, ECSCW '07, pages 21–40. Springer London.

Miller, D. R. H., Leek, T., and Schwartz, R. M. (1999). A Hidden Markov Model Information Retrieval System. In *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 214–221, New York, NY, USA. ACM.

Minka, T. and Lafferty, J. (2002). Expectation-propagation for the Generative Aspect Model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 352–359, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mizzaro, S. (1998). How Many Relevances in Information Retrieval? *Interacting with Computers*, 10(3):303–320.

Morris, M. R. (2007). Interfaces for Collaborative Exploratory Web Search: Motivations and Directions for Multi-user Designs. In *Proceedings of the ACM SIGCHI 2007 Workshop on Exploratory Search and HCI: Designing and Evaluating Interfaces to Support Exploratory Search Interaction*. `http://research.microsoft.com/en-us/um/people/merrie/papers/merrie_exploratory_search_wkshop_camera_ready.pdf` (retrieved 2015-09-16).

Morris, M. R. and Horvitz, E. (2007). SearchTogether: An Interface for Collaborative Web Search. In *Proceedings of the 20th Annual Symposium on User Interface Software and Technology*, UIST '07, pages 3–12, New York, NY, USA. ACM.

Morris, M. R., Teevan, J., and Panovich, K. (2010a). A Comparison of Information Seeking Using Search Engines and Social Networks. In *Proceedings of the 4th International Conference on Weblogs and Social Media*, ICWSM '10, pages 291–294. Association for the Advancement of Artificial Intelligence (AAAI).

Morris, M. R., Teevan, J., and Panovich, K. (2010b). What Do People Ask Their Social Networks, and Why? A Survey Study of Status Message Q&A Behavior. In

Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, pages 1739–1748. ACM.

Mrvar, A. (2015). Pajek / Pajek-XXL versions 3.** and 4.**. `http://mrvar.fdv.uni-lj.si/pajek/` (retrieved 2015-07-29).

Murakami, H., Mitsuhashi, K., and Senba, K. (2012). Creating User's Knowledge Space from Various Information Usages to Support Human Recollection. *Lecture Notes in Computer Science*, 7345:596–605.

Murphy, K. P. (2012). *Machine Learning A Probabilistic Perspective*. The MIT Press.

Myers, J. L., Well, A. D., and Lorch, R. F. (2010). *Research Design and Statistical Analysis*. Routledge.

Nagpal, A., Hangal, S., Joyee, R. R., and Lam, M. S. (2012). Friends, Romans, Countrymen: Lend Me Your URLs. Using Social Chatter to Personalize Web Search. In *Proceedings of the Conference on Computer Supported Cooperative Work*, CSCW '12, pages 461–470, New York, NY, USA. ACM.

Nayyar, A. (2015). Estimate the Dissemination of Social and Mobile Search for Different Information Needs Using Websites as Proxies. Master's thesis, Technische Universität München. Supervised by Christoph Fuchs and Georg Groh.

Neef, D., Siesfeld, A. G., and Siesfeld, T. (1998). *The Economic Impact of Knowledge*. Butterworth Heinemann.

Newby, G. (1996). Metric Multidimensional Information Space. In *Proceedings of Text Retrieval Conference*, TREC. `http://www.petascale.org/papers/newby-trec5.pdf` (retrieved 2016-01-17).

Noll, M. G. and Meinel, C. (2007). Web Search Personalization via Social Bookmarking and Tagging. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, ISWC/ASWC 2007, pages 367–380. Springer Berlin Heidelberg.

Norušis, M. J. (2011). *IBM SPSS Statistics 19 Advanced Statistical Procedures Companion*. Addison Wesley Pub Co Inc.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563.

Nurmi, P. and Bhattacharya, S. (2008). *Identifying Meaningful Places: The Non-parametric Way*, pages 111–127. Springer Berlin Heidelberg.

Nussbaumer, R. (2014). Classes of Information Needs and Derived Implications for the Search Process, Especially Considering Social and Mobile Aspects. Bachelor's thesis, Technische Universität München. Supervised by Christoph Fuchs and Georg Groh.

Oeldorf-Hirsch, A., Hecht, B., Morris, M. R., Teevan, J., and Gergle, D. (2014). To Search or to Ask: The Routing of Information Needs Between Traditional Search Engines and Social Networks. In *Proceedings of the 17th Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 16–27, New York, NY, USA. ACM.

Olguin, D. O., Gloor, P. A., and Pentland, A. (2009). Capturing Individual and Group Behavior with Wearable Sensors. *AAAI Spring Symposium on Human Behavior Modeling*.

Olguin, D. O. and Pentland, A. (2009). Sensible Organizations: A Sensor-Based System for Organizational Design and Engineering. In *Proceedings of International Workshop on Organizational Design and Engineering*, IWODE'09.

Osborne, J. and Waters, E. (2002). Four Assumptions of Multiple Regression that Researchers Should Always Test. *Practical Assessment, Research & Evaluation*, 8(2).

Pan, W., Aharony, N., and Pentland, A. (2011). Composite Social Network for Predicting Mobile Apps Installation. *AAAI Conference on Artificial Intelligence*.

Panovich, K., Miller, R., and Karger, D. (2012). Tie Strength in Question & Answer on Social Network Sites. In *Proceedings of the Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1057–1066. ACM.

Pirolli, P. (2009). An Elementary Social Information Foraging Model. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 605–614.

Pirolli, P. and Card, S. (1995). Information Foraging in Information Access Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 51–58.

Powers, D. M. (2011). Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Qiu, F. and Cho, J. (2006). Automatic Identification of User Interest for Personalized Search. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 727–736, New York, NY, USA. ACM.

Raftopoulou, P., Tryfonopoulos, C., Petrakis, E. G., and Zevlis, N. (2013). DS4: Introducing Semantic Friendship in Distributed Social Networks. In *Proceedings of the 21st International Conference on Cooperative Information Systems*, CoopIS'13, pages 185–203.

Rao, A., Spasojevic, N., Li, Z., and DSouza, T. (2015). Klout Score: Measuring Influence Across Multiple Social Networks. `http://arxiv.org/pdf/1510.08487.pdf` (retrieved 2016-02-29).

Ricci, F., Rokach, L., and Shapira, B. (2015). *Recommender Systems Handbook*. Springer US, second edition.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Duxbury Advanced Series.

Rivest, R. L., Shamir, A., and Adleman, L. (1978). A Method for Obtaining Digital Signatures and Public-key Cryptosystems. *Communications of the ACM*, 21(2):120–126.

Rogers, E. M. (1983). *Diffusion of Innovations*. Free Press, 3rd edition edition.

Ronald, S. (1998). More Distance Functions for Order-based Encodings. In *Proceedings of the International Conference on Evolutionary Computation Proceedings*, pages 558 – 563. IEEE.

Rousseau, F. and Vazirgiannis, M. (2013). Graph-of-Word and TW-IDF: New Approach to Ad Hoc IR. In *Proceedings of the 22nd International Conference on Information & Knowledge Management*, CIKM '13, pages 59–68, New York, NY, USA. ACM.

Runkler, T. A. (2015). *Data Mining: Modelle und Algorithmen intelligenter Datenanalyse*, volume 2. Springer Vieweg.

Saracevic, T. (2007a). Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance. *Journal of the American Society for Information Science and Technology*, 58(13):1915–1933.

Saracevic, T. (2007b). Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144.

Saramäki, J. and Onnela, J.-P. (2007). Structure and Tie Strengths in Mobile Communication Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7332–7336.

Schäfer, T. (2011). *Statistik II – Inferenzstatistik*. VS Verlag.

Schedl, M., Hauger, D., and Schnitzer, D. (2012). A Model for Serendipitous Music Retrieval. In *Proceedings of the 2nd Workshop on Context-awareness in Retrieval and Recommendation*, CaRR '12, pages 10–13, New York, NY, USA. ACM.

Schilit, B., Adams, N., and Want, R. (1994). Context-Aware Computing Applications. In *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*, pages 85–90.

Schmidt, K.-U., Sarnow, T., and Stojanovic, L. (2009). Socially Filtered Web Search: An Approach Using Social Bookmarking Tags to Personalize Web Search. In *Proceedings of the Symposium on Applied Computing*, SAC '09, pages 670–674, New York, NY, USA. ACM.

Semmler, G. (2013). Influence Models auf Facebook Daten. Bachelor's thesis, Technische Universität München. Supervised by Jan Hauffa and Georg Groh.

Serdyukov, P. and Hiemstra, D. (2008). Modeling Documents as Mixtures of Persons for Expert Finding. *Lecture Notes in Computer Science*, 4956:309–320.

Shadbolt, N., Hall, W., and Berners-Lee, T. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3):96 – 101.

Shah, C. and Pomerantz, J. (2010). Evaluating and Predicting Answer Quality in Community QA. In *Proceedings of the 33rd International SIGIR conference on Research and Development in Information Retrieval*, SIGIR '10, pages 411–418. ACM.

Shapira, B. and Zabar, B. (2011). Personalized Search: Integrating Collaboration and Social Networks. *Journal of the American Society for Information Science and Technology*, 62(1):146–160.

Shapiro, S. S. and Wilk, M. B. (1965). An Analysis of Variance Test for Normality (for complete samples). *Biometrika*, 52(3/4):591–611.

Shokouhi, M. (2007). Central-Rank-Based Collection Selection in Uncooperative Distributed Information Retrieval. *Lecture Notes in Computer Science*, 4425:160–172.

Shokouhi, M. and Si, L. (2011). Federated Search. *Foundations and Trends in Information Retrieval*, 5(1):1–102.

Si, L. and Callan, J. (2003). Relevant Document Distribution Estimation Method for Resource Selection. In *Proceedings of the 26th Annual International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '03, pages 298–305. ACM.

Si, L., Jin, R., Callan, J., and Ogilvie, P. (2002). A Language Modeling Framework for Resource Selection and Results Merging. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*.

Smirnova, E. and Balog, K. (2011). A User-Oriented Model for Expert Finding. *Lecture Notes in Computer Science*, 6611:580–592.

Smyth, B., Coyle, M., and Briggs, P. (2012). HeyStaks: A Real-World Deployment of Social Search. In *Proceedings of the 6th Conference on Recommender Systems*, RecSys '12, pages 289–292. ACM.

Sohn, T., Li, K. A., Griswold, W. G., and Hollan, J. D. (2008). A Diary Study of Mobile Information Needs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 433–442. ACM.

Spink, A., Jansen, B. J., Wolfram, D., and Saracevic, T. (2002). From E-Sex to E-Commerce: Web Search Changes. *Computer*, 35(3).

Steichen, B., Ashman, H., and Wade, V. (2012). A Comparative Survey of Personalised Information Retrieval and Adaptive Hypermedia Techniques. *Information Processing & Management*, 48(4):698–724.

Streeter, L. A. and Lochbaum, K. E. (1988). Who Knows: A System Based on Automatic Representation of Semantic Structure. *Recherche d'Information Assistée par Ordinateur*, pages 380–388.

Szulanski, G. (1996). Exploring Internal Stickiness: Impediments to the Transfer of Best Practice Within the Firm. *Strategic Management Journal*, 17:27–43.

Tajfel, H. (1970). Experiments in Intergroup Discrimination. *Scientific American*, 223:96–102.

Tajfel, H. and Turner, J. (1979). An Integrative Theory of Intergroup Conflict. In *The Social Psychology of Intergroup Relations*, pages 33–47. Monterey, CA: Brooks/Cole.

Tang, J., Chang, Y., and Liu, H. (2014). Mining Social Media with Social Theories: A Survey. *SIGKDD Explorations Newsletter*, 15(2):20–29.

Tang, M.-C., Ting, P.-H., and Sie, Y.-J. (2012). Exploring Evaluation Criteria of Social Navigational Tools on Social Media: A Case Study of aNobii. In *Proceedings of the 4th Information Interaction in Context Symposium*, pages 21–26. ACM.

Teevan, J., Morris, M. R., and Bush, S. (2009). Discovering and Using Groups to Improve Personalized Search. In *Proceedings of the 2nd International Conference on Web Search and Data Mining*, WSDM '09, pages 15–24, New York, NY, USA. ACM.

Teevan, J., Morris, M. R., and Panovich, K. (2011). Factors Affecting Response Quantity, Quality, and Speed for Questions Asked via Social Network Status Messages. In *Proceedings of the 5th International Conference on Weblogs and Social Media*, ICWSM '11, pages 630–633. Association for the Advancement of Artificial Intelligence (AAAI).

Thudt, A., Hinrichs, U., and Carpendale, S. (2012). The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries Through Information Visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1461–1470, New York, NY, USA. ACM.

Tigelaar, A. S., Hiemstra, D., and Trieschnigg, D. (2012). Peer-to-Peer Information Retrieval: An Overview. *Transactions on Information Systems (TOIS)*, 30(2).

Tinati, R., Carr, L., Hall, W., and Bentwood, J. (2012). Identifying Communicator Roles in Twitter. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 1161–1168, New York, NY, USA. ACM.

Turtle, H. and Croft, W. B. (1990). Inference Networks for Document Retrieval. In *Proceedings of the 13th Annual International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '90, pages 1–24, New York, NY, USA. ACM.

Turtle, H. and Croft, W. B. (1991). Evaluation of an Inference Network-based Retrieval Model. *Transactions on Information Systems (TOIS)*, 9(3):187–222.

UCLA: Statistical Consulting Group (2014). R Data Analysis Examples: Ordinal Logistic Regression. `http://www.ats.ucla.edu/stat/r/dae/ologit.htm` (retrieved 2015-11-25).

Uzun, A., Salem, M., and Kupper, A. (2013). Semantic Positioning – An Innovative Approach for Providing Location-Based Services Based on the Web of Data. In *Proceedings of 7th International Conference on Semantic Computing*, ICSC '13, pages 268–273.

Uzzi, B. (1997). Social Structure and Competition in Interfirm Networks: The paradox of Embeddedness. *Administrative Science Quarterly*, 42:35–67.

Vallet, D., Cantador, I., and Jose, J. M. (2010). Personalizing Web Search with Folksonomy-Based User and Document Profiles. In *Proceedings of 32nd European Conference on Information Retrieval*, ECIR '10.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths.

Voigt, C. (2015). Information Retrieval using Topic Models. Master's thesis, Technische Universität München. Supervised by Christoph Fuchs and Georg Groh.

Vuorikari, R. and Koper, R. (2009). Ecology of Social Search for Learning Resources. *Campus-Wide Information Systems*, 26(4):272–286.

Wagner, C., Liao, V., Pirolli, P., Nelson, L., and Strohmaier, M. (2012). It's not in his tweets: Modeling topical expertise of Twitter users. In *Proceedings of the IEEE International Conference on Social Computing*, pages 91 – 100.

Wang, C., Paisley, J., and Blei, D. M. (2011). Online Variational Inference for the Hierarchical Dirichlet Process. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, AISTATS '11.

Wang, X. H., Zhang, D. Q., Gu, T., and Pung, H. K. (2004). Ontology Based Context Modeling and Reasoning Using OWL. In *Proceedings of the 2nd Annual Conference on Pervasive Computing and Communications Workshops*. IEEE.

Wasserman, L. A. (2005). *All of Statistics – A Concise Course in Statistical Inference*. Springer.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Watts, D. J., Dodds, P. S., and Newman, M. E. J. (2002). Identity and Search in Social Networks. *Science*, 296(5571):1302–1305.

Wei, X. (2007). *Topic Models in Information Retrieval*. PhD thesis, University of Massachusetts Amherst.

Wei, X. and Croft, W. B. (2006). LDA-based Document Models for Ad-hoc Retrieval. In *Proceedings of the 29th Annual International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 178–185, New York, NY, USA. ACM.

White, R. W., Kules, B., Drucker, S., and Schraefel, M. (2006). Supporting Exploratory Search: A Special Section of the Communications of the ACM. *Communications of the ACM*, 49(4):37–39.

Workman, T. E., Fiszman, M., Rindflesch, T. C., and Nahl, D. (2014). Framing Serendipitous Information-Seeking Behavior for Facilitating Literature-Based Discovery: A Proposed Model. *Journal of the Association for Information Science and Technology*, 65(3):501–512.

World Wide Web Consortium (2013). W3C Semantic Web Activity. `http://www.w3.org/2001/sw/` (retrieved 2015-11-20).

Xu, J. and Croft, W. B. (1999). Cluster-based Language Models for Distributed Retrieval. In *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 254–261. ACM.

Yang, K.-H., Chen, C.-Y., Lee, H.-M., and Ho, J.-M. (2008). EFS: Expert Finding System based on Wikipedia link pattern analysis. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 631 – 635.

Yang, Z., Zhong, S., and Wright, R. N. (2005). Anonymity-preserving Data Collection. In *Proceedings of the 11th SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 334–343, New York, NY, USA. ACM.

Yimam, D. (1996). Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR Approach. In *ECSCW Workshop: Beyond Knowledge Management: Managing Expertise*, ECSCW. ACM.

Zhai, C. and Lafferty, J. (2001). A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of the 24th Annual International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 334–342, New York, NY, USA. ACM.

Zhang, J. and Ackerman, M. S. (2005). Searching for Expertise in Social Networks: A Simulation of Potential Strategies. In *Proceedings of the 2005 International SIGGROUP Conference on Supporting Group Work*, GROUP '05, pages 71–80, New York, NY, USA. ACM.

Zhang, X., Cole, M., and Belkin, N. (2011). Predicting Users' Domain Knowledge from Search Behaviors. In *Proceedings of the 34th International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1225–1226. ACM.

Zhao, X., Salehi, N., Naranjit, S., Alwaalan, S., Voida, S., and Cosley, D. (2013). The Many Faces of Facebook: Experiencing Social Media as Performance, Exhibition, and Personal Archive. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1–10. ACM.

Zhu, X., Ming, Z.-Y., Zhu, X., and Chua, T.-S. (2013). Topic Hierarchy Construction for the Organization of Multi-source User Generated Contents. In *Proceedings of the 36th International SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 233–242, New York, NY, USA. ACM.

THANK YOU!

Numerous people supported the creation of this thesis. First and foremost, I would like to express my sincere gratitude to my advisor, Priv.-Doz. Dr. Georg Groh, who laid the foundation for this special topic, provided valuable advice, and supported me throughout the whole process. I also would like to thank my second advisor, Prof. Dr. Helmut Krcmar for interesting and helpful discussions.

I really enjoyed being part of the "Applied Informatics – Cooperative Systems" department of Prof. Dr. Johann Schlichter. Many ideas used in this thesis have been discussed and improved during the discussions with the team (Prof. Dr. Johann Schlichter, Dr. Florian Schulze, Dr. Michele Brocco, Hanna Schäfer, Jan Hauffa, Dr. Wolfgang Wörndl, Claudius Hauptmann, Dr. Niklas Klügel and many others).

Furthermore, I want to thank all the marvelous students I had the honor to work with (Oriana Baldizan, Cordt Voigt, Carola Boettcher, Ruth Nussbaumer, Akash Nayyar, Jessica Bauer, Christina Kopp, Nissrine El Marchoum, Sahar Taissa, Trinh Viet Doan, Belinda Zahra, and many others). The main experiment of the thesis was made possible through support of the students in the *Social Computing* and *Social Gaming* courses held during summer term 2015 at TU München – a large "thank you" to each and every one of you for your trust and collaboration!

I also would like to thank my family for support, encouragement, and motivation and in particular those who directly witnessed my highs and lows on this journey: my wife Jacqueline and my sons Marc & Jan – thank you for your (nearly) endless understanding, support and sincere love!