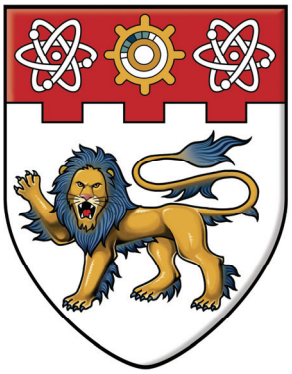

NANYANG TECHNOLOGICAL UNIVERSITY

EPISTEMOLOGY-BASED SOCIAL SEARCH FOR EXPLORATORY INFORMATION SEEKING

MAO YUQING

SCHOOL OF COMPUTER ENGINEERING

2012

EPISISTEMOLOGY-BASE SOCIAL SEARCH FOR EXPLORATORY INFORMATION SEEKING

MAO YUQING

2012

Epistemology-based Social Search for Exploratory Information Seeking

Mao Yuqing

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2012

Acknowledgements

I would like to express my immense gratitude to my supervisors, Dr. Shen Haifeng and Prof. Sun Chengzheng. They have given me great intellectual freedom to pursue my interests and have provided patient guidance and kind help throughout my PhD study. Their advice and encouragement have been absolutely essential for the completion of this thesis. From them, I have not only learned presentation, writing and technical skills, but also learned to appreciate high quality in research.

I received what seemed like endless help from members of the Parallel & Distributed Computing Center in many capacities. They have provided me with all the convenience I need for my study and research. Thanks also to the staff and technicians in Flinders University for their support and participation when I was attending the oversea research attachment in Australia.

I would like to extend my thanks to Fan Hongfei, Agustina, Xu Yi, and all members of the Advanced Collaborative Technology Research Group, past and present, for their friendship, support and interest in my work. I also thank my friends at NTU and Flinders University, Hao Jianan, Jin Jiangming, Li Yusen, Li Zengxiang, Liu Cheng, Song Liang, Bill Chen, and many others. It is them that make my journey of research a pleasant one.

I owe a great deal to my parents who gave a life and unstinting help to me. I am eternally grateful to them for all the opportunities they made available to me, and for the love and support they have given me along the way.

Finally, I am deeply indebted to my dear wife Ying for her tremendous support for my oversea study. Very special thanks to my son, Jihang. Of course, he is always my anchor that keeps me going. Words cannot truly express how much I owe them. I would never accomplish this thesis without their love and encouragement.

Table of Contents

List of Figures

List of Tables

Abstract

As the Internet and communication technologies become ever more pervasive, we see an astounding number of new information seeking behaviors. People often start with some vague information need and iteratively seek and select bits of information that cause the data needs and behavior to evolve over time. This phenomenon is referred to as exploratory information seeking, which is open-ended, persistent, and multifaceted. In general, formulating proper keywords and evaluating search results are common difficulties in exploratory information seeking. One possible solution is to utilize social cues provided by a large number of users.

The main objective of this thesis is to improve exploratory information seeking by social search, i.e., utilizing the wisdom of crowds. The main contribution of this thesis research is an epistemology-based social search framework, where search epistemologies – aggregated and well-structured information packages derived from successful search processes contributed by numerous searchers – are effectively shared, reused, and refined by others with same or similar search interests. This framework contains several distinctive components that are well-organized and work in cooperation with each other.

The first component is the epistemology generation component, where the search epistemologies can be automatically derived from successful search processes of massive users doing exploratory information seeking. The generative process of an epistemology is modeled using a probabilistic topic model with social tags. An approach for query reformulation and results ranking is proposed based on this model.

The second component is the epistemology search component, where the search epistemologies can be retrieved for users doing exploratory information seeking with diverse requirements. A social-interest-based query suggestion and results diversification approach is proposed to support exploratory information seeking, while the social interest is discovered by employing the kernel principal component analysis on the related queries and results.

The third component is the epistemology editing and refining component, where the search epistemologies can be collaboratively edited in a consumer-led interactive search process. An information provision on demand approach is proposed to help information consumers acquire non-existent information, while invited information providers from relevant social networks jointly refine the epistemologies to meet the consumer's needs.

The fourth component is the epistemology services component, which makes the social epistemology-based search systems viable, reliable, and sustainable. A trust model for trust inference through credit flow in the epistemology-based social search network is proposed, which allows personalized measures derived from trust propagation to be naturally established on the objective ground. A non-monetary incentive mechanism for encouraging users to contribute their epistemologies is proposed, where a knowledge bartering process can be automated through the online silk road that maximizes the social welfare within a social search community. Furthermore, experimental results based on the prototype system *Baijia* demonstrate these methods can enhance the epistemology-based social search and improve exploratory information seeking with better performance.

Glossary

CCP	- Cycle Covering Problem
EIS	- Exploratory Information Seeking
EPISOSE	- Epistemology- based Social Search for Exploratory Information Seeking
ILP	- Integer Linear Programming
IPOD	- Information Provision On Demand
IR	- Information Retrieval
KPCA	- Kernel Principal Component Analysis
LDA	- Latent Dirichlet Allocation
MAP	- Mean Average Precision
MRR	- Mean Reciprocal Rank
NDCG	- Normalized Discounted Cumulative Gain
OWL	- Web Ontology Language
QA	- Question Answering
RDF	- Resource Description Framework
RMT	- Real-Money Trading

ROC curve	- Receiver Operating Characteristic curve
RWR	- Random Walk with Restart
SE	- Search Engine
SEO	- Search Engine Optimization
TF-IDF	- Term Frequency – Inverse Document Frequency
TN	- True Negative
TP	- True Positive
TREC	- Text Retrieval Conference
SVM	- Support Vector Machine
UGC	- User-Generated Content
VOD	- Video on Demand

Chapter 1

Introduction

1.1 Overview

Nowadays the World Wide Web (the Web) has become the main source of information, not only because of the abundance of information, but also because of the convenience of retrieving information. The Web has a higher degree of freedom than conventional media since it is possible for anyone to publish any information at any time from any place. More importantly, search engines (SE) such as *Google* and *Bing* make it easy for people to retrieve information on the Web. Web search is constantly getting more powerful and easy to use [190], pushing people's expectation increasingly higher as they expect web search to deliver any information they want even when they have difficulties to clearly translate their information needs to precise queries. However, the reality is otherwise as current search engines are generally incompetent for exploratory search tasks whose information needs are vague and hard to be described by queries and that often take place over time.

This thesis presents my research work on effectively supporting exploratory information seeking (EIS) tasks through an epistemology-based social search approach, where users with the same or similar search goals or interests can help each other by sharing their personal knowledge or search processes.

1.1.1 Background

The Web has become indispensable in many people's daily lives. The size, heterogeneity and dynamic nature of the Web make search engines very different from traditional information retrieval systems. In traditional information retrieval (IR), a user's information need is usually clear and straightforward. Schneiderman et al. [147] defined information need as "the perceived need for information that leads to someone using an information retrieval system in the first place." However, the Web has extended significantly both the range of people's search goals and the range of resources available [4]. A query to a search engine can hardly be treated as a one-time conception of the user's information need. Broder [28] investigated "the need behind the query" and found out that only about 20% of the queries were intended to find a specific web site that the user had in mind, whereas others were about a topic or intended to "perform some web-mediated activity". Furthermore, the information need can change during a search session. Through the interaction with a search engine, the user who started with a vague information need can gradually increase their knowledge in the task domain from the retrieved results and can consequently refine queries iteratively to make them close to the actual information need.

A variety of terms have been used to describe an exploratory information seeking activity: exploratory search, informational search, interactive search, human-computer information retrieval. The following definition of exploratory search has been widely accepted in this research area:

"Exploratory search can be used to describe an information-seeking problem context that is open-ended, persistent, and multi-faceted; and to describe information-seeking processes that are opportunistic, iterative, and multi-tactical. In the first sense, exploratory search is commonly used in scientific discovery, learning, and decision-making contexts. In the second sense, exploratory tactics are used in all manner of information seeking and reflect seeker preferences and experience as much as the goal" [103].

For the sake of explanation, we make this EIS process more concrete with the following example:

Under the threat of swine or bird flu, people may want to seek useful information on the Web for protecting their families. Because they know little about the flu or medicine in general, they have difficulties in formulating proper keywords and therefore what they normally do is through trial and error: "symptom", "epidemic", "virus", "vaccination", "immunity", and so on.

Furthermore, given a set of trial keywords, a search engine may return considerable number of results about swine or bird flu. People with little medical knowledge have difficulties in evaluating the information quality and therefore what they normally do is by means of exhaustive comparison: viewing as many results as possible and picking up relatively good ones out of the mass.

Information seeking in the proposed research domain is subtly different from classical IR. In IR, the target is typically known, its existence is confirmed prior to query issuance, and the user's task is to create a well-formed query that will retrieve relevant documents at

the top of a ranked list. In contrast, in information seeking, it is uncertain whether the information to be sought exists or whether the user is able to find it. Major SEs such as *Google* and *Bing* can handle classical IR well, given the significant advancement in ranking algorithms and instant answers (e.g., weather forecasts or stock quotes), but encounter great difficulties in handling activities associated with exploratory information seeking, which are far beyond single-session lookup tasks.

1.1.2 The Social Search Approach to Exploratory Information Seeking

The past few years have witnessed the transformation of the resource-centric one-to-many Web 1.0 media, in which professional designers publish static documents for the benefit of an unseen passive audience, to the people-centric many-to-many participatory Web 2.0 platforms, where people can actively network - share, communicate, collaborate, and interact - with anyone, anywhere, anytime. As such, the abundance of information on the Web has changed the way people seek information.

In Web 1.0, Web pages were limited in quantity, content, and source (most by government agencies, organizations, and large corporations); therefore, people used to do navigational searches just to find a specific Web site or page, such as the Web site of the *World Health Organization* [28]. In Web 2.0, information on the Web has never been so abundant in quantity, rich in content, and diverse in source that a search process tends to be exploratory [103] in the sense that it is non-trivial to precisely formulate a query to express the search goal or clearly evaluate the search results retrieved by the query against the search goal, especially for an information seeker who is not an expert in the domain of the search task.

The evolution of Web from 1.0 to 2.0 has transformed Web search from navigational to exploratory, but current SEs are lagging behind the demand for effectively supporting EIS tasks. With a current SE (e.g., *Google*), one could only perform an EIS task in a rather ad hoc way: generate numerous queries with bunches of ambiguous keyword combinations, evaluate a pile of search results for each query according to their relevance to the overall search goal, and manually take down the relevant results in a document so that the search process can continue in the next session at a different time.

While information seekers can easily handle most of their navigational searches alone, it is often beyond their capability to perform a complex EIS task in isolation, especially when they are not experts in the subject domain being searched. "*The Wisdom of Crowds*" [162], where large groups of cognitively and socially diverse individuals have proven superior to the elite few in solving problems, fostering innovation, or decision making, is the rationale behind the emerging technique of social search. This type of search technique empowers seekers to help each other find information online by sharing their domain knowledge, search efforts or search results and can be used to address the key EIS issues, including query formation, result evaluation, and outcome documentation and dissemination.

1.2 Research Objective and Framework

1.2.1 Research Objective

While different social search techniques address different EIS issues largely in isolation, such as keywords, results, or certain type of meta-information about results [36], the

objective of this research is to contribute a holistic epistemology-based social search solution that can systematically and consistently address all the major exploratory search issues.

1. Query formation: existing queries in a social epistemology inform a new user of what good queries were formulated by others and what else they can build on.
2. Result evaluation: existing Web page URLs in a social epistemology inform a new user of what good search results were short-listed by others and how they were evaluated by others through re-rankings, ratings, annotations, and comments.
3. Outcome documentation and dissemination: the outcome of a user's EIS process is documented into a new epistemology or an existing social epistemology and disseminated as a social epistemology.

1.2.2 Research Framework

Figure 1.1 shows the framework of this thesis research. This work is built on the foundation of the "search epistemology" concept and the EPISOSE (Epistemology-based Social Search for Exploratory Information Seeking) solution framework. The search epistemologies are aggregated and well-structured information packages derived from successful search processes, such as queries, results, rankings, annotations, comments, and inquiries acquired by numerous seekers in information collection, pre-processing, filtering and post-processing. The term "epistemology" in philosophy is concerned with the nature and scope of knowledge, not only the knowledge itself, but also why people

know the knowledge and how the knowledge is acquired. We adopt this term as we also care about the knowledge in a whole search process rather than only search results. The EPISOSE framework leverages existing algorithmic SEs to collect and pre-process information and human knowledge to filter and post-process information.

A prototype system *Baijia* is designed and implemented under the guidance of the EPISOSE framework, where the EIS can be effectively supported by allowing multiple users to share, reuse, and refine the intimate search epistemologies contributed by others in a social search network with same or similar search interests.

Epistemologies are automatically generated by integrating social tagging into a generative probabilistic model for query reformulation and context-aware result ranking in EIS. Social intents are discovered from multiple concepts of users' behavior in EIS to suggest queries and retrieve epistemologies. Synchronous and asynchronous cooperation and collaboration in EIS is enhanced by supporting collaborative epistemology editing in the social search system.

Social networking is incorporated in the social search system, by strengthening its role in forming trust-aware social search networks and acquiring trustworthy information within the social networks to improve users' search experience. An innovative incentive mechanism is developed to encourage users to contribute and refine epistemologies. This solution is evaluated by conducting novel user-based and simulation-based experiments.

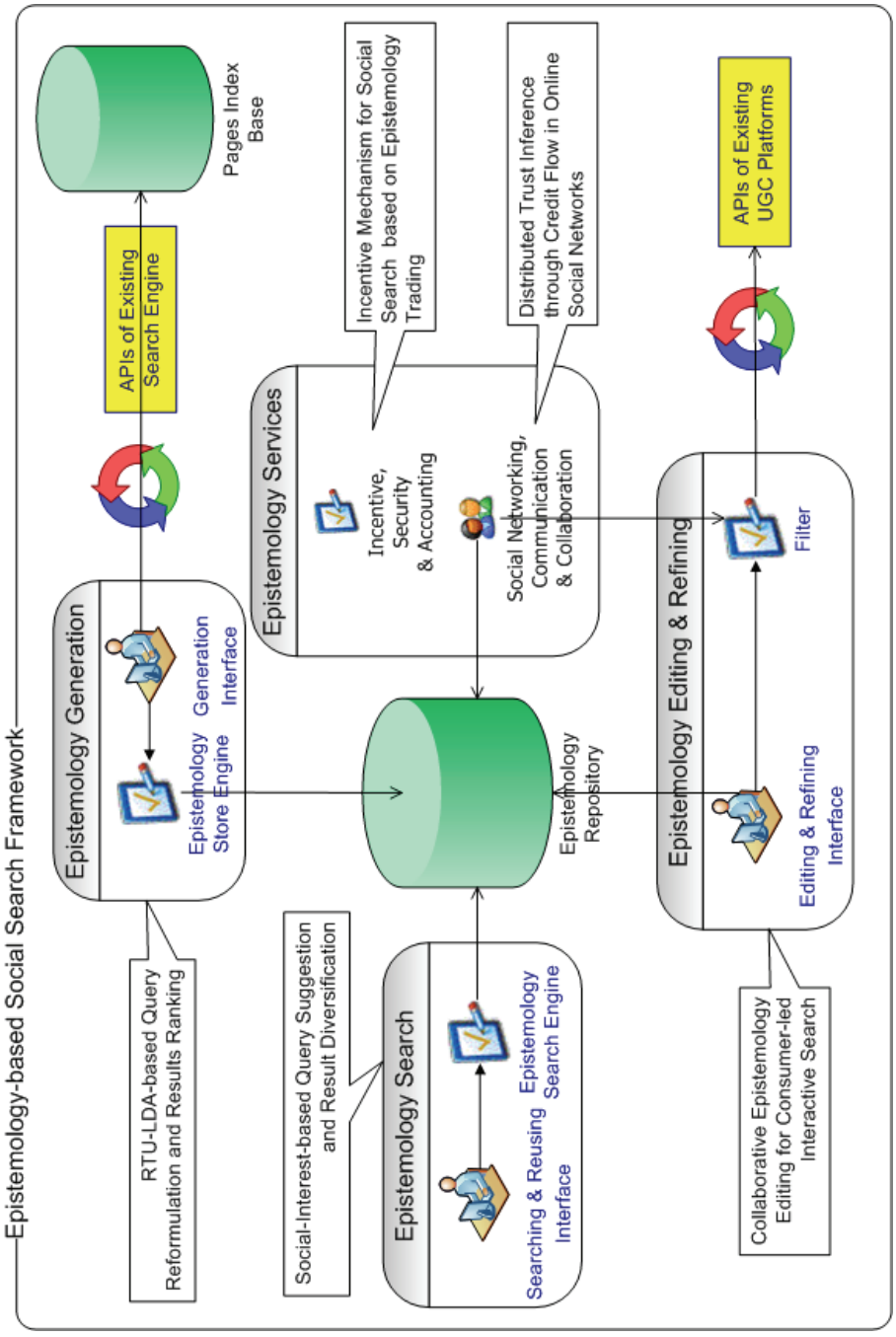

Figure 1.1 Research Framework

1.3 Major Contributions

This thesis investigates a new frontier in the web search domain – EIS, which focuses on understanding and supporting searches that may result from ill-defined information needs, require explorative search strategies, or have personal development as a primary objective. We propose an innovative solution to supporting effective EIS by epistemology-based social search, where epistemologies contributed by numerous seekers are aggregated and well-structured information packages derived from search processes, such as queries, results, rankings, annotations, comments, and inquiries.

We have developed a number of novel techniques in our social epistemology research, including a probabilistic topic model for epistemology generation, a social-interest-directed diversification approach for epistemology retrieval, an information provision on demand approach for epistemology editing, a credit-flow-based approach to trust management in epistemology-based social search and a trading-based incentive mechanism for epistemology sharing. Although these techniques were developed for epistemology-based social search in this thesis, we expect they can generally benefit research in information retrieval, web mining, social networking, and human computer interaction. Major contributions of this thesis include:

1. An epistemology-based social search framework EPISOSE, which makes the best of both conventional SEs' immense power of information collection and pre-processing and human users' knowledge of information filtering and post-

processing. This framework can be applied to the design and implementation of a range of social search systems with different strategies and algorithms.

2. A new query reformulation approach based on a novel probabilistic topic model RTU-LDA to discovering the latent semantic relationships between the queries and the URLs. It can not only discover related queries that cannot be clustered by existing query clustering approaches but also rank retrieved results according to the similarities of probability distributions over the latent topics among the queries and the URLs.
3. A social-interest-directed technique that integrates social interest mined from query logs using a probabilistic model on query-URL bipartite graphs. The social interest, discovered through kernel principle component analysis on the related queries and URLs with random walk on the bipartite graph, can suggest highly diverse queries that are yet closely related to a given query, and retrieve relevant and diverse results for that query at the same time.
4. An information provision on demand approach (IPOD) that can help users acquire non-existent information. This approach builds and exploits the social network of likeminded users in EPISOSE framework. Information is provided on demand through a consumer-led interactive search process where invited information providers from relevant social networks jointly create content on the fly to meet the consumer's needs.
5. A credit-flow-based trust model CoreTrust that allows personalized trust measures to be naturally established on the objective ground. The key concept of this model is to infer trust by tracing the credit flow within a social network,

where the trust between a pair of nodes can be derived from the credit flowing from one node into the other and the relative risk disparity between them. This model is inspired by the physical and mathematical properties and the power flow study in electrical grids.

6. A trading-based incentive mechanism Silk Road that encourages users to contribute and share epistemology, where one possessing a piece of knowledge in a field can directly exchange it for another piece of knowledge in a different field perceived to have equivalent value. By maximizing the total of weights for the supply-demand pairs in the knowledge market, this mechanism can achieve the maximum social welfare in social search networks.

1.4 Organization of the Thesis

This thesis studies the issues in EIS and proposes a holistic epistemology-based social search solution that can systematically and consistently address these issues. This thesis is organized according to the EPISOSE solution framework (Figure 1.1). This chapter has introduced EIS, social search, the motivation of our research, and major contributions of this thesis. The rest of the thesis is organized as follows:

Chapter 2 "Literature Review" provides some background knowledge and existing work related to social search.

Chapter 3 "The Epistemology-based Social Search Solution" describes the proposed solution to effective EIS, including the search epistemology concept, the EPISOSE framework, and the *Baijia* prototype system.

Chapters 4 - 8 present the key techniques in the solution framework. Each chapter is dedicated to a particular issue.

Chapter 4 "Epistemology Generation" studies the issue of deriving search epistemologies from successful search processes contributed by users. The automatic generative process of an epistemology is modeled using a probabilistic topic model with social tags. The model is used for query reformulation and results ranking.

Chapter 5 "Epistemology Search" proposes a social-interest-directed query suggestion and results diversification approach, supporting epistemology retrieval from users with diverse information needs. The social interest is discovered by employing the kernel principal component analysis on the related queries and results.

Chapter 6 "Epistemology Editing" discusses the IPOD approach that helps information consumers acquire non-existent information through a consumer-led interactive search process.

Chapter 7 "Trustworthy Social Networking" addresses the trust management issue in the epistemology services component. The CoreTrust model propagates trust through tracing credit flow in epistemology-based social search networks. This work is inspired by the physical and mathematical properties and the power flow study in electrical grids.

Chapter 8 "Non-monetary Incentive Mechanism" addresses the incentive issue of encouraging users to contribute and share epistemologies. An incentive mechanism based on epistemology trading is proposed for users to trade their knowledge without involving monetary means.

Chapter 9 "Conclusions and Future Work" concludes the thesis with a summary of major contributions and directions for future work.

Chapter 2

Literature Review

In this chapter, we review existing work related to various methods of improving user's web search experience: 1) web search, including the classic information retrieval techniques and the emerging web mining of previous searches (section 2.1), 2) social activities, including collaboration in search processes and online social networks (section 2.2), and 3) social search, an interdisciplinary approach combining web search and social activities (section 2.3).

2.1 Web Search

2.1.1 Information Retrieval

Classic IR concentrates on the storage and acquisition of electronic information in digital libraries. However, theories, techniques and tools that constitute the traditional approaches to the organization and processing of information may still be applicable in Web search, such as IR formal models, ranking schemes, and algorithms.

Developing formal models for IR has a long history. A model of information retrieval "predicts and explains what a user will find relevant given the user query" [67]. Most IR models, including Boolean model, vector space model, and probabilistic model, are

firmly grounded in mathematics and thus provide a formalization of ideas that can be implemented (see details in Appendix A).

Web IR is different from classical IR because there are several issues that do not exist in classical IR circles, such as modeling of the Web graph (including graph theory, random graphs, scale-free and small-world networks), link analysis, crawling, low quality content and quality issues, heterogeneous users (not just librarians), etc. In applications of link analysis to the Web, including PageRank [121] and HITS [82], the links between Web pages are basically used as a factor in computing the importance of a Web page or document for ranking Web search results. For example, the PageRank method is an important component of the Google Web IR engine. The PageRank score of a page *i* depends on the PageRank scores of pages pointing to it and on the number of links going from the pages. It is defined as:

$$P(i) = (1-d)\frac{1}{\#\,pages} + d \sum_{j|j\ links\ to\ i} \frac{P(j)}{\#\,links}$$

PageRank aims at returning high quality documents (i.e., documents from trusted sites), rather than only returning documents that closely match the query terms (i.e., by using any of the classic IR models). The rationale behind PageRank is that Web pages that are linked from many places are probably worth looking at, and they are likely high quality pages. In this sense, PageRank is a primitive form of social search, as it assesses the importance of a page according to the votes for that page by all other pages on the Web. However, it utilizes only the static collective opinion of the Webmaster community,

while current social search applications are to elicit the opinions of people in specific search processes.

2.1.2 Web Mining

To SEs, the ultimate goal is to understand the goals of information consumers and to return what they want. Unfortunately, the processing of natural languages and the deployment of semantic web still has a long way to go. Researchers are currently focusing on mining the web content/structure and users' interaction (web usage) in order to rank the results pertinent to a search query. Web mining is the application of data mining techniques to web-based data for the purpose of learning or extracting knowledge. Web mining encompasses various branches of knowledge discovery techniques, e.g., finding natural groupings of pages by clustering (through unsupervised learning), or assigning a query to a predefined category by classification (through supervised learning).

In this section we are concerned with Web usage mining, which is useful for learning about a web system's users. The idea of mining and reusing others' successful searches has emerged with the rise of SEs. A typical example is keyword suggestions in SE such as real time Google Suggest, where users will see the keyword suggestions when they type theirs in the search box. More advanced studies have been carried out to discover users' key concepts in search processes by mining the query log data. For instance, Wang & Zhai [173] proposed new methods for mining term association patterns from search logs to support query refinement in a more effective way. Query expansion is often used to enhance and modify the queries in SEs' logs for reusing [22] [119]. For example, Armin [6] proposed a new query expansion method for iterative reformulations of the

queries to build collaborative IR systems. Lee [91] attempted to generate recommendations for users with similar queries by uncovering patterns in users' queries and subsequent browsing with algorithms of data mining.

Considering that explicit feedback would interrupt users' natural searching behaviors [44], most work is based on mining implicit feedback from users, e.g., the click-through data. For instance, Radlinski & Joachims [130] proposed an improved ranking function learned from preference data using a ranking SVM (Support vector machine). Baeza-Yates & Tiberi [12] attempted to extract semantic relations from query logs using the query-click bipartite graph. Recommender systems in the IR field [113][146] are based on the collaborative filtering technique, which has been successfully applied in Amazon.com[1], where users will see "people who liked this product also like that product" when they browse a product's web page [95]. Some web search systems try to collect users' profiles together with the queries of users and measure their similarity for personalized web search [76]. Herlocker, et al. [65] proposed an algorithmic framework for performing collaborative filtering based on users' profile, and Koren et al. [85] presented a probabilistic framework for faceted metadata and collaborative user relevance model to customize personalized search. I-SPY [157] provided personalized search results for a particular community of users by capturing relationships between queries and results pages for particular users, and can help new searchers by capitalizing on past successful searches of similar users in a community. Porqpine [129] used an automatic query

[1] http://www.amazon.com

expansion model based on user profile and a collaborative filtering mechanism to re-rank search results.

Although research in Web usage mining reuses others' searches, it is not regarded as genuine social search because users are not explicitly engaged in reusing processes [23][185][165], which instead are purely based on reusing optimal queries by analyzing query logs and search results from SEs [16][77]. Since the objective is to make the reusing process transparent when users are doing search, the reusing is done in the background by the SEs. Implicit reusing of others' searches is based on the view that information seeking is an individual activity. In the following sections, we shall discuss and challenge this view.

2.2 Social Activities

2.2.1 Collaborative User Experience

The following factors are relevant to the success of a web search: 1) to rank search results according to the relevance to the search goals (by the SEs), 2) to provide information that can be easily accessed (by information providers), and 3) to formulate precise keywords that express the search goals (by information consumers). In the previous section, we have reviewed the efforts made on the SE side. However, in recent years, social computing, Web 2.0 applications and other techniques have impacted a new collaborative user experience, both for information providers and consumers.

2.2.1.1 User-Generated Content

The amount of information provided on the Web has skyrocketed since the inception of Web 2.0, mainly owing to the ever-growing user-generated content (UGC) published in social media exemplified by uncountable wikis, blogs, microblogs, social networking newsfeeds, and social annotations. "Sharing" is the cornerstone of many social websites such as bookmark sharing (Delicious[2]), photo sharing (Flickr[3]), and video sharing (YouTube[4]). A large number of people are interested in particular information and are encouraged to describe it or annotate it (they may tag freely to organize their own content retrieval or to make "shared" easy for others to retrieve).

To information providers, some research on search engine optimization (SEO) [132] has been done to achieve easy accessibility of their information. However, the abuse of SEO can cause the information overload by SEs. To make IR more efficient, information can be published on the web as documents accompanied by semantic markup [19]. Semantic Web, which is the spirit of Web 3.0, may have the potential to support EIS [60]. Semantic publication is intended to provide a way for computers to understand the structure and even the meaning of the published information, using semantic web languages like RDF (Resource Description Framework) and OWL (Web Ontology Language).

Ontology is usually developed for a specific information domain, which is then used to formally represent the data in such domain. The complexity of deep ontologies has led

---

2 http://delicious.com
3 http://www.flickr.com
4 http://www.youtube.com

some researchers to eschew ontologies in favor of a different approach [118]. One of the most important formats of UGC - social tagging - is a development generating considerable interest at the moment. Social tags arise with the rapid development of social media websites, where users create or upload content (resources), annotate it with freely chosen words (tags), and share these annotations with others. These websites are known as collaborative tagging systems and have been coined a name “folksonomy", which is a combination of “folk" and “taxonomy" [156]. Compared to ontologies which attempt to define parts of the data world more carefully and to allow mappings and interactions between data held in different formats [98], social tags serve for a very different purpose. They represent a structure that emerges organically when individuals manage their own information requirements. Users can assign keywords to documents or other information sources, rather than a centralized form of classification,

Folksonomy has attracted increasing attention in Web document classification and search[13][75]. For example, Yanbe et al. [192] proposed a search model combining standard link-based ranking method with the one using data from social bookmarking. Zhou et al. [202] proposed a generative model for social annotation tags and provided methods of combining language models with it for IR. However, tags are often different from keywords submitted to an SE and need to be improved for facilitating search [66].

2.2.1.2 Collaboration on the Web

To information consumers, field knowledge and search skills are essential to a successful information seeking task. As individual users’ knowledge and skills are limited, collaboration is a common practice in search activities in real world. However, as

navigating and seeking for information on the Web are generally regarded as a single user activity, both Web browsers and search engines are mostly designed with single-user interfaces and functionalities. In recent years, researchers have rethought about this view. Hansen and Jarvelin [62] have discovered that collaborative activities are an important characteristic of information seeking and retrieval processes in professional task-based Web search. The survey conducted by Spence et al. [161] has shown that the lack of expertise is the primary reason of collaborative information seeking. Another survey from Morris [114] has revealed that a large proportion of users engaged in collaborative activities when doing web search.

In response to these findings, some works have implemented several search tools to explicitly support collaborative search. Collaborative search, a special paradigm of explicit social search, is the process of more than one person jointly performing the same search task. Tools that facilitate collaborative search can be generally classified as collaborative Web browsers, collaborative search tools and collaborative search systems.

Collaborative Web browsers exemplified by the *W4 browser* [49] and *GroupWeb* [55], allow a group of distributed "slave" users to synchronously view the same Web page navigated by the "master" user. Romano N. et al. [137] integrated information retrieval with group support systems to build a collaborative IR environment. Users perform only standard, single-user searches, but can add comments to pages they find. These comments are then visible to other members of the group who visit these same pages later. A group history can be viewed, which lists the pages visited by the entire group.

Collaborative search tools, such as *SearchTogether* [115] and *HeyStaks* [159], are emerging to support general-purpose EIS tasks through multi-user communication, collaboration, and interaction. For example, *SearchTogether* allows a group of distributed users to collaborate synchronously or asynchronously on the same search task, where query formation is addressed by query awareness that provides per-user query histories in order to help users maintain awareness of others search strategies, the result evaluation is addressed by division of labor that best leverages each member's knowledge/skill set, and the outcome dissemination is addressed by an instant messenger that disseminates search results among group members.

Collaborative search systems also enable collaboration among small groups of people for shared web search tasks. For example, the system implemented by Pickens et al. [126] provides an algorithmic mediation to enhance search and communication activities among a team. CIRR [198] provided a system task console with a "Group Report" of relevance judgments, search queries and search results. Físchlár-DT [155] realized supporting collaborative search for Video on a tabletop interface. These works do not focus on reusing others' contributions but on taking the advantage of a pre-defined group of users to collaboratively perform a common web search task.

Collaborative search distinguishes itself from other social search techniques by using a group of pre-selected collaborators to perform the same search task, whereas users involved in social searches are generally anonymous and do not perform the same search task. Research efforts on collaborative search have been largely independent of those on social search. In contrast, in the epistemology-based social search solution, collaborative

search functionality has been seamlessly integrated with other social search functionalities to support EIS.

2.2.2 Social Networking

Over the past several years, online social networking web sites have become increasingly popular. They offer services for making friends like *MySpace*[5], *Facebook*[6] and *Google+*[7], and for writing blogs like *LiveJournal*[8] and *BlogSpot*[9]. These sites have exploded in extreme popularity among web users: *MySpace* claims to have over 100 million users, and has been observed to receive more page hits than *Google* [117], while *Facebook* has grown to more than 800 million active users since it was opened to the general public.

Due to the incredible amount of media coverage, many people might believe that social networks are a recent invention [54]. In fact, social network analysis has been a multidisciplinary field from the very beginning [83], which can be traced back to three original disciplines: Sociology, Psychology and Mathematics.

According to Knoke's research in sociology [83], actors and relation are the two main elements of any social network. Actors may be individual persons or collectivities such as informal groups and formal organizations. Common examples of an individual actor include an employee in a corporate work team, or a student attending a high school

---

5 http://www.myspace.com
6 http://www.facebook.com
7 http://plus.google.com
8 http://www.livejournal.com
9 http://www.blogspot.com

graduation. Collective actors might be firms competing in an industry, or political parties holding seats in parliament. A relation is generally defined as a specific kind of contact, connection, or tie between a pair of actors. Relations may be either directed, where one actor initiates and the second actor receives, or non-directed, where mutuality occurs. A relation is not an attribute of one actor but a joint property that exists only as long as both actors maintain their association.

Research in psychology suggests that social relations are critically important for collaboration. In e-learning settings, the continued participation depends on the social interaction with professors and other students, even more than the course content and material [104]. Recruitment to political and social movements often depends on pre-existing social relations [5][106]. Studies of data taken from other online groups show that individuals with a strong sense of attachment to a group are more likely to cooperate[84], and closely connected groups are more supportive of members [179].

In traditional social network analysis, graph-theoretic algorithms are useful for mathematically determining the derived facts about entities in the network. For example, one of the most popular algorithms in network analysis is the computation of "centrality degree" for a node, which is measured by adding the number of incoming links on a node, and provides some indication of how important that node might be [175]. Analysis of group membership within a social network can indicate likely membership relations [87]. Input to a link discovery tool can be pre-filtered to identify individuals who have varying levels of contact with known threats, cutting down analysis time and providing a small boost to accuracy [182].

2.3 Social Search

As social activities on the Web are attracting growing interest, social search has been increasingly adopted to address those difficulties in EIS by utilizing the wisdom of crowds [162] in recent years. Regarding that individual users' knowledge and skills are limited, social search has transformed a search process from a solitary activity to a collaborative one.

There is no a widely accepted definition of social search. Generally speaking, searches carrying with human-labor are collectively called "social search", as opposed to algorithm-based searches. The core idea of social search is combining human intelligence with computer algorithms, and the central tenet of social search is "People + Algorithms > Algorithms". It focuses on how social groups can influence and potentially enhance the ability of algorithms to find meaningful information for end users, namely "better search through people" [4].

Social search takes many forms, and there have been some commercial social SEs in the market, such as *Mahalo*, *Yoople*, and *Scour*, which allow information consumers to re-rank search results through voting and editing and to share their re-ranked results. Recently Question Answering (henceforth QA) has been a very popular form of information seeking, which allows a user to seek help from others regarding a specific question. Representative QA systems include *Yahoo! Answers*[10] and *Baidu Zhidao*[11]. QA

[10] http:// answers.yahoo.com

has become popular and the TREC (Text Retrieval Conference) workshop comprises a track of QA too. QA is a noticeable method for a wide range of information needs. For example, Bian et al. [21] presented an effective social media search method to retrieve factual information from QA archives. A few commercial systems use human labor to simulate the ability to process natural-language-style search queries, such as *ChaCha*[12], which offers live assistance from experts to guide ordinary users and suggest URLs to visit in performing web search tasks.

At the same time, social search provides the opportunity to access the *Hidden Web* or *Deep Web*, which embraces the pages not directly indexable by conventional SEs but has a size about 2-50 times larger than the *Visible Web* or *Surface Web* [34], i.e., the part of the Web directly indexable by SEs. For example, deep web documents that are excluded from the crawling for commercial or technical reasons could be identified as unpublished human knowledge by users of social search, and made available for other users. Therefore, with a wide scope and enormous benefits, social search is a promising direction to improve the user experience in EIS.

---

11 http://zhidao.baidu.com

12 http://www.chacha.com

Chapter 3

The Epistemology-based Social Search Solution

The most important characteristic of our epistemology-based social search solution is the contribution and reusing of search epistemologies to support EIS. To reach this target, we proposed an Epistemology-based Social Search framework (EPISOSE). Containing a collection of components with different functions, it can be a comprehensive and systematic solution for EIS. The EPISOSE framework can be applied to the design and implementation of a range of social search systems with different strategies and algorithms. To validate the feasibility and effectiveness of the framework, we have designed and implemented a prototype system *Baijia* with the guidance of the framework and conducted a set of experiments to measure the system's performances in supporting EIS. In this chapter, we first introduce the fundamental element "epistemology" of our solution, then illustrate the constitution of the EPISOSE framework, and describe the details of the *Baijia* system, finally present the results of our experiments.

3.1 Introduction

There are two major reasons why people cannot effectively share their searches. One is that there is no proper structure to effectively organize whatever that should be shared. The other is that there is no proper interface for users to effectively contribute or retrieve. We propose the concept of "epistemology" to structure Web search

experience/knowledge and provide novel interfaces for users to easily contribute and retrieve epistemologies. As shown in Figure 3.1, an epistemology is a hierarchically structured knowledge unit derived from a seeker's successful EIS process, including a set of queries $q_1,\ldots,q_n$ (n ≥ 1) formulated by the seeker, a list of Web page URLs $u_1,\ldots,u_m$ (m ≥ 1) clicked by the seeker for each query, and the seeker's re-ranking, rating (e.g. using a number of stars), annotations (e.g. using tags), and comments of each of these URLs.

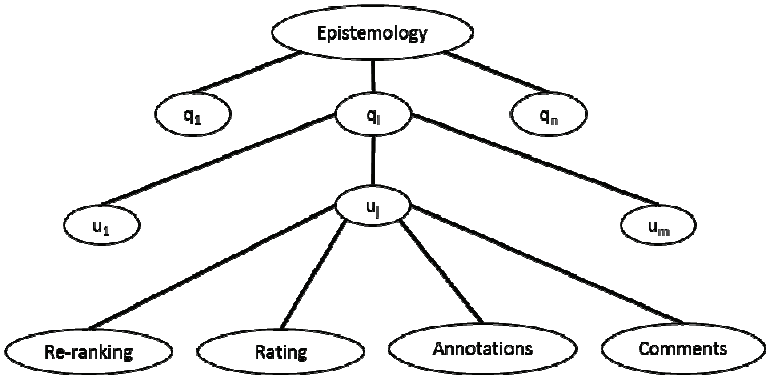

Figure 3.1 Epistemology structure

An example of the RDF description of epistemology is shown in Figure 3.2.

```
<Epistemology rdf:about="http://baijia.com/epistemology/2009">
   <dc:title>Bird Flu Preparation</dc:title>
   <dc:description>Search for bird flu symptoms and preparation</dc:description>
   <queries rdf:resource="http://baijia.com/epistemology/query/2015"/>
   <queries rdf:resource="http://baijia.com/epistemology/query/2016"/>
...
   <comments rdf:resource="http://baijia.com/epistemology/comment/2027"/>
...
   <evaluations rdf:resource="http://baijia.com/epistemology/evaluation/2028"/>
...
   <foaf:creator>
      <foaf:Person  rdf:about="http://baijia.com/epistemology/user/Hui Shi">
      </foaf:Person>
   </foaf:creator>
   <foaf:editor>
      <foaf:Person  rdf:about="http://baijia.com/epistemology/user/Han Fei">
      </foaf:Person>
   </foaf:creator>
...
</Epistemology>
```

Figure 3.2 An example of epistemology structure

An epistemology becomes social when it is shared with others who may use it if it adequately meets their search goals or otherwise edit it by contributing their knowledge, e.g. their re-rankings, ratings, annotations, and comments of the existing URLs, additional new relevant URLs, and additional new queries. Therefore, a social epistemology is an aggregate knowledge unit derived from numerous EIS processes contributed by many users who share the same or similar search goals or interests. A social epistemology can be expressed based on the RDF, as shown in Figure 3.3.

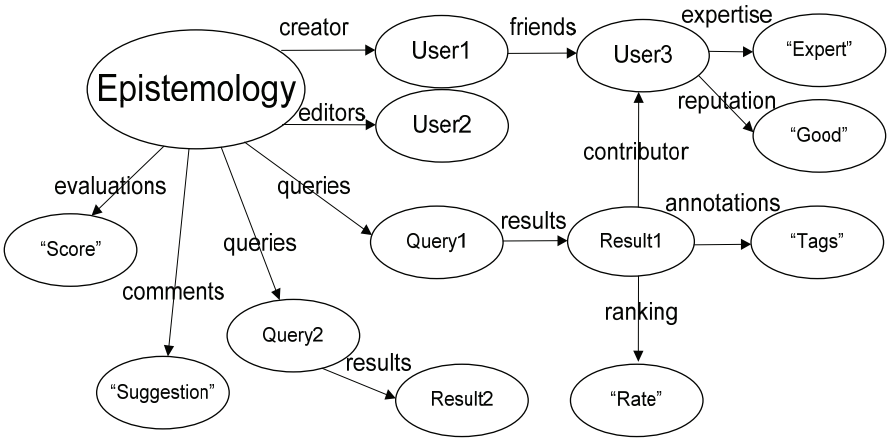

Figure 3.3 A graph representation of the social epistemology

Social epistemology holistically addresses all the key EIS issues and seamlessly integrates synchronous or asynchronous collaborative search activities into a social search process.

- Query formation: existing queries in a social epistemology inform a new seeker of what good queries were formulated by others and what else they can build on.
- Result evaluation: existing Web page URLs in a social epistemology inform a new seeker of what good search results were short-listed by others and how they

were evaluated by others through re-rankings, ratings, annotations, and comments.

- Outcome documentation and dissemination: the outcome of a seeker's EIS process is documented into a new epistemology or an existing social epistemology and disseminated as a social epistemology.
- Seamless integration of collaborative search: within an EIS process, the seeker can establish a collaborative search session to invite other seekers to jointly create a new or edit an existing social epistemology at any time.

3.2 The Epistemology-based Social Search Framework

The EPISOSE framework is the cornerstone and kernel of our solution to improve the efficiency of EIS. This research is organized around the framework, since it is novel and uniquely different from previous works as highlighted in the following parts. We start with the illustration of the framework architecture. Then we introduce the components constructing the framework. Further discussions of the comparison of the EPISOSE with other Web search frameworks can be found in Appendix B.

3.2.1 Architecture of the Framework

Contrary to the clear distinction between information providers and information consumers in current SEs, where providers are never engaged in a consumer's search process, the social epistemology technology takes a novel "prosumer" [163] approach that blurs the distinction between providers and consumers in that consumers are also

providers to their peers in social search. The positive prosumer feedback cycle is the key to addressing the exploratory nature of a search task. Nonetheless, the technology is not proposed to supersede current SEs; instead, it is a value-adding technology as it leverages the existing SE's power of gathering and ranking information infused with the social power of generating knowledge out of the gathered information. The architecture of EPISOSE is shown in Figure 3.4.

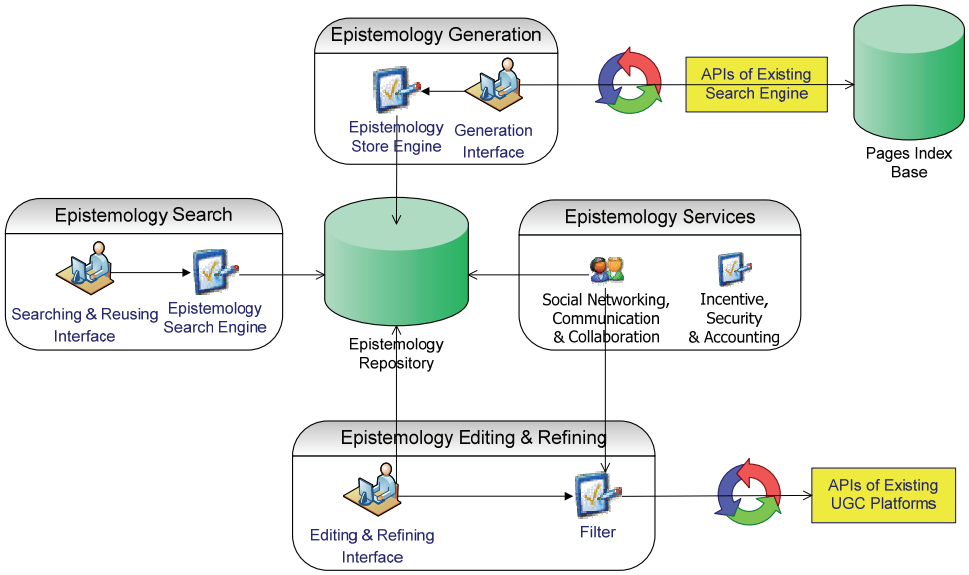

Figure 3.4 The EPISOSE Framework

3.2.2 Components of the Framework

As depicted in Figure 3.4, the EPISOSE framework consists of the following major components.

3.2.2.1 Epistemology Search

This component is for a social search community to share and reuse search epistemologies. While prosumers input search queries through the *Searching & Reusing Interface*, the *Epistemology Search Engine* will search the epistemology repository and

return the relevant search epistemologies. These epistemologies were contributed by other prosumers with the same or relevant search interests or goals through the *Generating Interface* powered by the *Epistemology Store Engine*. If no relevant epistemology is found, the *Search Engine* (e.g., Google) will search the *Pages Index Base* and return relevant pages according to the keywords. Prosumers can generate their own epistemologies from the result pages returned by the SE. In this component, knowledge discovery techniques (such as classification or clustering) can be applied to analyze the relevance of the input queries to the stored search epistemologies in the epistemology repository. Because each epistemology is tagged with plenty of additional information by processing raw information returned by an SE with the prosumer's intimate knowledge and understanding, the *Epistemology Search Engine* has a very high probability to return relevant epistemologies pertinent to a query. Consequently, other prosumers can save their time in repeating the course of collecting and processing raw information through an SE.

3.2.2.2 Epistemology Generation

This component helps prosumers conveniently generate their search epistemologies through the *Epistemology Generating Interface*, and systematically store them to the *Epistemology Repository* through the *Epistemology Storing Engine* to facilitate epistemology sharing and reusing. Search epistemology may include the sub-sequence of search keywords, the approbatory results selected by the prosumer, the commentaries on these results added by the prosumer, other useful information about the search topic provided by the prosumer, and evaluation of the search provided by other prosumers.

This component is significantly different from answering questions or adding annotations because stored search epistemologies can be easily retrieved by *Epistemology Searching & Reusing* without relying on SEs, which have difficulties in retrieving most relevant information if proper keywords are hard to be formulated. Since the epistemologies are generated in the whole EIS process, the retrieval of epistemologies would be a heuristic search. Prosumers can be quickly led to the final search goal by the epistemologies generated from others' successful searches.

3.2.2.3 Epistemology Editing and Refining

This component is dedicated to supporting collaborative search by allowing a group of prosumers to jointly edit and refine the same epistemology synchronously or asynchronously. Prosumers can refine others' relevant epistemologies retrieved from the epistemology repository, or discuss with the contributors of certain epistemologies in order to better refine them, or invite buddies in their social networks to join their ongoing search processes.

The EPISOSE framework can support communication and collaboration in a group of EIS, including: assigning each member different search tasks, dealing with epistemology refining conflicts between members, or transmitting directions from experts, and so on. As existing epistemologies can be refined by subsequent prosumers, they tend to be more relevant and accurate as they are refined by more prosumers. Hence, the advantage of EPISOSE is that it provides a method for prosumers to contribute to the social search community by refining the existing epistemologies. As a result, the epistemologies would

become more and more perfect and the framework is self-reinforcing as the number of searches increases.

3.2.2.4 Epistemology Repository

This component stores search epistemologies contributed by the prosumers in a social search community. Search epistemologies are information packages regarding specific search goals. They are indexed by the combination of the search goal and relevant keywords for the sake of easy retrieval by the *Epistemology Search Engine*. The *Epistemology Repository* also stores other relevant information about the epistemologies as the base of some advanced applications. For example, building social network needs the information about who contributes the epistemology, communicating between prosumers needs the information about where the epistemology is transmitted.

3.2.2.5 Epistemology Services

This component has the following functions:

Social Networking, Communication & Collaboration: this component has the following functions: *Social Networks Building* helps prosumers with the same or similar search goals build up social networks to complete search tasks together. While prosumers are doing EIS, they would be likely to look for help from others. Prosumers can find people with same hobbies or similar information requirements from search epistemologies and thence build a social network with them. Furthermore, the EPISOSE framework can adopt effective strategies to search for expertise in social networks, so that exploratory searches can benefit from the building of social networks. When one is not sure about what she/he is looking for, seeking advices from experts is always a good option.

Communication Facilities allow prosumers in a social search community to communicate via tools such as messenger or email. *Collaborative Session Management* allows a group of prosumers to share their search epistemologies in an ongoing search process synchronously or asynchronously. Therefore, collaborative search is seamlessly integrated into the social search approach in such a way that a prosumer can create a collaborative session on an epistemology they generated using *Epistemology Generation* or retrieved using *Epistemology Search* and then invite others from their social networks to join the session.

Services Management: this component provides some common services for making social search for EIS viable, reliable, and sustainable: *Incentive* encourages prosumers to share their epistemologies; *Security* handles issues related to privacy and security in a social search community; *Accounting* can estimate the prosumers' contribution in order to establish a revenue model for social search; and so on.

3.3 The *Baijia* Prototype System

The EPISOSE framework can be applied to the design and implementation of a range of EIS systems with different strategies and algorithms. With the guidance of the EPISOSE framework, we have designed and implemented a prototype system *Baijia* to validate our epistemology-based social search solution. In the following sections, we will present how a prosumer search community is enabled to share and reuse the social epistemologies in the system. A more detailed illustration is included in Appendix C.

3.3.1 Data Collection

We selected the AOL query logs [123] as the initial dataset for the *Baijia* system. The dataset includes {AnonID, Query, QueryTime, ItemRank, ClickURL}, where AnonID presents an anonymous user ID number, ClickURL is the URL user clicked and ItemRank is the rank of the clicked item on the listed results. The initial epistemology repository of *Baijia* is built by importing the dataset and the search epistemologies are automatically generated for each user. For some selected topics, search epistemologies are clustered by keywords of the queries, and formed by specified rules, e.g., different weights are given to URLs according to the number of click times for the same queries. With the initial epistemology repository, we can also setup experiments to evaluate the effectiveness of *Baijia* system by contrasting the new search processes with those in the query logs.

3.3.2 Overview

Figure 3.5 shows an epistemology window in the *Baijia* system. When the user submits a query to the system, the entry for the query will be added automatically to the epistemology, and the user can add selected pages for this entry by dragging them to the epistemology. The user can modify the epistemology by reorganizing the entries and adding appropriate descriptions. Further, the user may provide other references or own suggestions for the epistemology. While most previous work focused on providing more relevant pages for a specific query, *Baijia* can discover related queries in the epistemology repository. As such, our solution can better support EIS because users can

get the required information quickly and easily even they start with not-so-relevant queries.

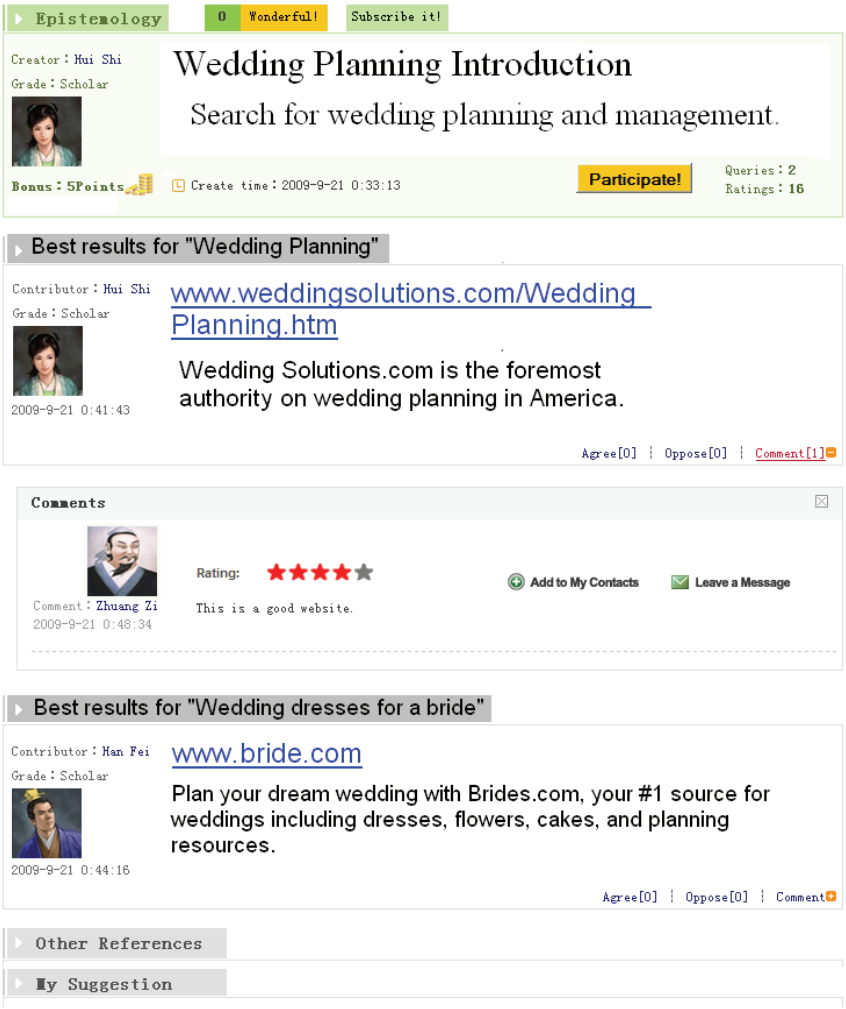

Figure 3.5 A window of epistemology generation

For an EIS process, the initial epistemology is automatically generated on the fly for the user by the system. The epistemology is private by default. Later on, the user may choose to refine and/or make it public. The relationship between queries is determined based on the user's Web browsing session; the selection of pages is based on the user's clicking actions.

Selected pages in an epistemology will be ranked and commented by other users, therefore the system will re-rank the pages in each epistemology dynamically. The ranking of a page is based on all received scores (one to five stars) for the page, and the reputation (honest or fraudulent) and expertise (newcomer or skilled) levels of each commenter.

One can choose to share or not to share one's search epistemologies. For the privacy's sake, a security mechanism is deployed in the system to allow users to share within a specific group, e.g., family members. It is helpful in some cases, e.g., colleagues searching for new techniques in order to get a project done. All members can take part in the search task and share epistemologies with each other. Furthermore, the system supports refinement of epistemologies and notification of updates of interested epistemologies. One can create an initial epistemology entry in the repository and then subscribe to it by clicking the "subscribe it!" button. The system will send a notification to the user when the epistemology is re-fined by others. One can subscribe (with proper authentication and authorization) to any epistemology no matter who has actually created it.

The epistemology search engine adjusts the weights between the epistemologies and the results from the SE according to the level of relevance between the queries and the epistemologies. The measurement of relevance is applied to the sorting of retrieved epistemologies. If no match of epistemology is found, results from the SE are given a higher priority, and the most relevant epistemologies are listed for the user's reference. Furthermore, as the computation is based on the queue of queries, the epistemologies are

re-ranked dynamically while the user continuously submits queries. Figure 3.6 shows a window of related epistemologies in *Baijia*.

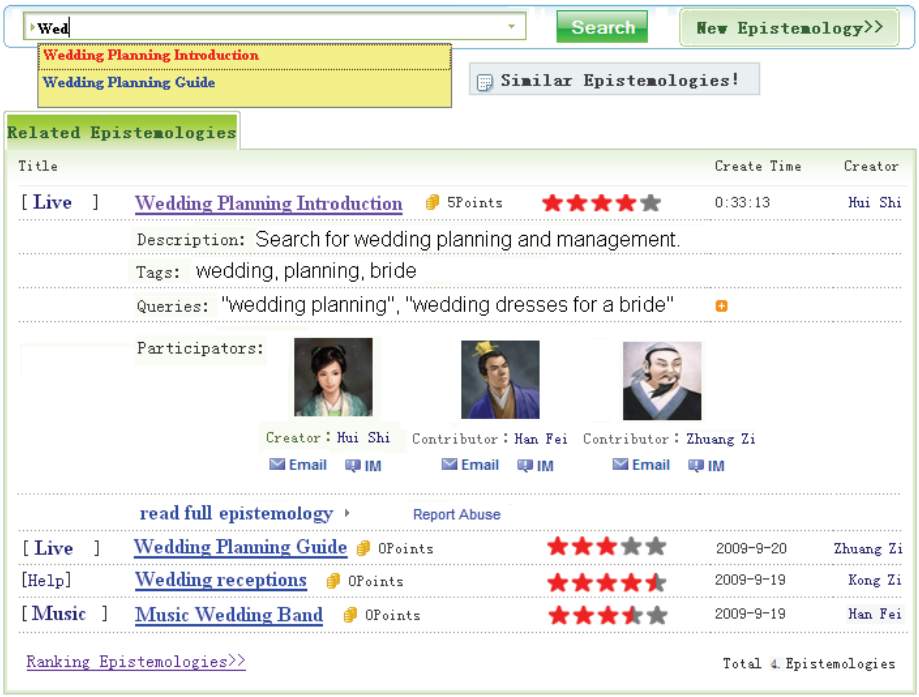

Figure 3.6 Epistemology Search

Moreover, a spam detection mechanism is in place to prevent contributions of epistemologies with malicious purposes. User can click the "Report Abuse" button to impeach the epistemology if it is suspected to be created for SEO purpose (e.g., epistemologies are ranked highly by illegitimate means for commercial purposes).

3.4 Preliminary Test

The main purpose of setting up experiments is to validate how much an epistemology-based social search system can outperform a conventional SE in supporting EIS.

3.4.1 Initialization

We first constructed the initial epistemology repository based on epistemologies automatically generated from the AOL query logs. The AOL query logs consist of about 20 million search records from about 650,000 users over three months. These records contain 10, 154, 742 unique queries and 1, 632, 789 unique URLs.

Although the dataset doesn't contain explicit users' feedback on search results, the URL clicking can be regarded as positive feedback because relative feedback signals generated from users' clicking behaviors have been proved to correspond well with explicit judgments [77]. Therefore it is possible to backtrack users' search processes according to the query logs. Reposing on this technical foundation, we used intelligent agents to simulate users' interactions with epistemologies and search results (based on AnonID).

In our experiments, search epistemologies are contributed and shared through the following steps:

Step 1:

"Users" completed their searches through iterative interaction with the system and contributed their search epistemologies. To simulate the contribution from users, we extracted every user's search processes from their queries. Each search process contains several queries that are contextually related. Cosine distance function is used to measure the contextual similarity between every two queries. We have totally extracted 1,201,497 search processes.

Step 2:

The system returned other users' search epistemologies that are relevant to the queries of the current user from its epistemology repository. To simulate the sharing of epistemologies, we retrieved the epistemology repository for relevant search epistemologies. An epistemology is relevant to a search process if its queries are similar to the search queries, and the selected pages of the epistemology completely/partially match the clicked URLs of the search.

Step 3:

If no relevant epistemology is found at step 2, the search process itself will be formulated as a search epistemology; otherwise, it will be integrated into existing relevant epistemologies. "Users" participated in the search activity by re-ranking the ranked results from the SE or the re-ranked results from other users. To simulate the refinement of epistemologies, a computer-generated score following a Gaussian distribution is assigned to every clicked URL to represent the judgment from the current user. Finally we have built 480,254 records in the epistemology repository.

Following the above steps, we have built up the initial epistemology repository for *Baijia* by importing all exploratory searches derived from the AOL query logs. It is worth pointing out that the initial epistemology repository can immediately benefit new exploratory information consumers, but the system can actually work without it. The system relies on the SE to build up the epistemology repository at its initial stage and gradually relies more on the epistemology repository.

3.4.2 Metrics

We adopt some metrics that have been widely used to evaluate the performance of SE, such as Mean Average Precision (MAP) [29], Precision at K (e.g., Precision@10) [29], and Normalized Discounted Cumulative Gain (NDCG) [74], to compare the performance of our epistemology-based social search system *Baijia* with that of the AOL SE. In particular, we introduce the following metric to *Baijia* in order to describe that the system is self-reinforcing as the number of searches increases.

EAR:

Epistemology Acquisition Rate (EAR) is the ratio of users' searches that can successfully retrieve relevant epistemologies. This metric is introduced to measure how many searches can benefit from the epistemologies in the repository. Obviously, the EAR of the AOL SE is always 0, since users' searches are not shared at all. In *Baijia*, we envisage that the more exploratory searches were performed, the higher EAR would be. We also use EAR together with other metrics, e.g., MAP scores, to compare the performance of our epistemology-based system with that of the AOL SE.

MAP:

Average precision of a query is defined as follows:

$$AP = \frac{1}{rel}\sum_{r=1}^{rel}\frac{pos_r}{r} \qquad (1),$$

where *rel* is the total number of documents relevant to the query, and pos_r is the position of the r^{th} relevant document in the list of all resultant documents. MAP is defined as the

mean AP over all queries. It can stably reflect the overall performance of a search system [171].

ISE:

We specially introduce the metric of Interactive Search Entropy (ISE) to measure the performance of a search system in supporting EIS in terms of the total number of queries issued in an EIS process. It represents the interaction times between users and the system; the average ISE of a search system can reflect how fast users can get the requested information from the search system. While searching with a conventional SE, users have to analyze the previous results and formulate next queries by themselves. In contrast, *Baijia* can accelerate a search process by taking advantage of search epistemologies in the repository. We anticipate the average ISE of *Baijia* would be significantly lower than that of the AOL SE.

NDCG:

Given a query, NDCG at position k is defined as:

$$NDCG(k) = \frac{1}{Z_k} \sum_{p=1}^{k} \frac{2^{S(p)} - 1}{\log(1+p)} \qquad (2),$$

where k is a particular top rank position, $S(p)$ is the score for rank p, denoting the graded relevance of the result at position p, and Z_k is a normalization factor derived from a perfect ranking algorithm, so that an ideal NDCG at position k will be equal to 1.0 and all NDCG calculations are relative values on the interval from 0.0 to 1.0. The NDCG values for all queries can be averaged to obtain a performance measure of a search system. In *Baijia*, users can give different scores to the same re-ranked search results from others.

Therefore we anticipate such an epistemology refinement mechanism would significantly improve the NDCG scores of *Baijia*.

With the metrics above, we are able to evaluate the overall performance of the *Baijia* system and test the conjecture that the system can provide higher precision, shorter search time, and better quality of search results owing to the contribution, reuse, and refinement of search epistemologies.

3.4.3 Experimental Results

We traced the generation of the epistemology repository. Our major concern is whether an exploratory search can benefit from the system's epistemology repository. Therefore, we computed the EAR, ISE, MAP and NDCG scores of the *Baijia* system at different stages. In our experiments, the MAP is computed as the mean of every exploratory search's AP, which is the weighted mean of precisions of all queries it consists of.

$$MAP = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{Q_i} w_j AP_j\right) \qquad (3),$$

where N is the total number of exploratory searches, Q_i is the total number of queries in the i^{th} exploratory search, AP_j is the AP of the j^{th} query in the i^{th} exploratory search and w_j is the weight of it according to its importance to the exploratory search (e.g. a query with more clicked URLs will be assigned a heavier weight than a query without any clicked URL).

The EAR value of *Baijia* is shown in Table 3.1, which increases as more exploratory searches are imported. Thus it can be seen that *Baijia* is self-reinforcing because the more

searchers contribute to the epistemology repository, the more likely new searchers will reuse previous epistemologies.

Table 3.1 Epistemology repository size and EAR at different stages

Number of exploratory searches imported	Epistemology repository size	EAR
20,000	7,612	18.35%
200,000	74,634	29.67%
400,000	151,392	34.20%
600,000	230,273	37.37%
800,000	311,167	39.93%
1,000,000	394,266	41.71%
1,201,497	480,254	42.52%

Figure 3.7 shows the MAP scores of the *Baijia* system as compared to those of the AOL SE (the original data). The results show that increase of the number of exploratory searches imported leads to improvement of MAP scores in the *Baijia* system while the MAP scores of AOL SE are steady. While more exploratory searches are imported, a user who submits an exploratory search will have a higher probability to get relevant search epistemologies from the epistemology repository. Since search epistemology is extracted from clicked URLs and unclicked URLs have been filtered out, the MAP scores of the *Baijia* system are clearly higher than those of the AOL SE, which can never benefit from previous searches at all. Furthermore, when search epistemologies in the epistemology repository are re-ranked according to users' feedback rather than random weights, the MAP scores will even be significantly improved.

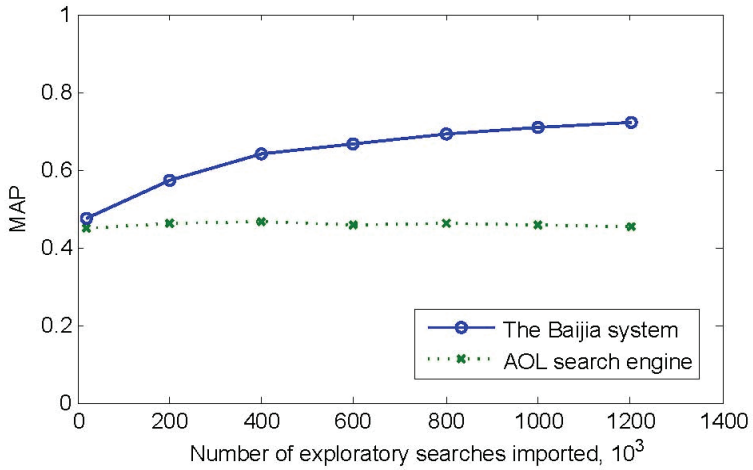

Figure 3.7 MAP scores of the *Baijia* system and the AOL SE

Unlike previous studies which mainly focus on augmenting search results with relevant data aggregated from the Semantic Web by pre-defined ontologies [29], our work aims at automatically constructing ontologies based on the sequence of queries in every EIS process.

In Table 3.2, we illustrate the average ISE for both *Baijia* and AOL. As about 26.3% of all exploratory searches were unsuccessful in the dataset (without clicking any URLs), and about 25.9% of all successful searches are one-step searches (such a search task requires no exploration at all), they should be excluded for the comparison study. After excluding unsuccessful searches and one-step searches, we can see a noticeable improvement of average ISE.

Table 3.2 ISE of the *Baijia* system and AOL SE

ISE	The *Baijia* System	AOL SE
Overall	1.5563	2.3734
Exclusive of unsuccessful searches	2.1924	3.7080
Exclusive of unsuccessful and one-step searches	3.0331	5.2076

Figure 3.8 shows NDCG@10 of *Baijia* and AOL. It can be seen that *Baijia* has obviously achieved a better ranking performance than that of AOL. This is because *Baijia* re-ranks the search results according to users' judgments, and the re-ranked results are more approximate to those of the perfect algorithm (ranking based on part of users' evaluation).

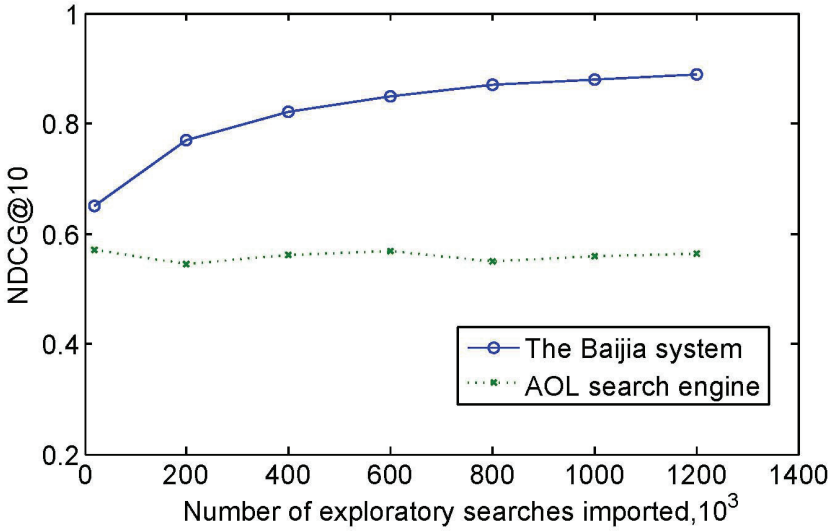

Figure 3.8 NDCG@10 of the *Baijia* system and the AOL SE

Efficiency is very important for the system performance of a search engine, where the time and space costs have a great influence on its usability. In the preliminary test, the *Baijia* system is rather inefficient, comparing to the commercial AOL search engine. However, there are approaches that can only decrease a little accuracy which give a very substantial increase in efficiency (e.g., to discard an amount of rarely-used queries). Further, in most EIS scenarios, a lot of time consumed in pre-processing and post-processing is a more serious problem than the efficiency of document searching. The efficiency of the whole search process in *Baijia* is improved since the user might avoid using a number of queries to search, and looking at each document in the results of the search engine.

3.5 Summary

Conventional SEs are incompetent in the situations where the users have difficulties in formulating proper keywords and must struggle to evaluate search results. In this chapter, we have proposed a novel epistemology-based social search framework EPISOSE for supporting EIS, where search epistemologies – aggregated and well-structured information packages derived from successful search processes contributed by a mass of prosumers – are effectively shared, reused, and refined by others with same or relevant search interests or goals. We have also implemented a prototype system *Baijia* based on the framework and conducted a set of experiments to prove that the proposed solution can outperform a conventional SE in supporting EIS, and the richer the social epistemology repository, the shorter the search process and the higher the quality of sought information. Preliminary usage study indicates that utilizing SEs' immense power and human prosumers' intelligence is an effective and pragmatic solution to exploratory web search.

We have introduced *Baijia* on our intranet to get some initial usage feedback. Most feedback confirms the improvement of the search efficiency in various situations. A main dissatisfaction is that no enough well-refined up-to-date search epistemologies were available to benefit from at the elementary stage. This situation will be improved as the user number increases and the epistemology repository grows. As EPISOSE is an epistemology-based social search framework, *Baijia* relies more on SEs at its initial stage and is self-adaptive to the growing human factors of the system.

Chapter 4

Epistemology Generation

The objective of this thesis is to propose an epistemology-based social search solution to holistically addressing the key issues in EIS including query formation, result evaluation, and outcome documentation and dissemination. In the previous section, we have introduced the general structure of the EPISOSE framework. In the following chapters, we will introduce some critical issues we have addressed in the components of this framework. This chapter describes a novel probabilistic topic model with social tags we have proposed for the epistemology generation issue. First, Section 4.1 gives the background of extracting user interests for the aggregation of exploratory searches. Then, some related works are provided in Section 4.2. The topic model and its application to epistemology generation are described in Section 4.3. Section 4.4 discusses the experimental methodology and results.

4.1 Introduction

Epistemologies could be either manually generated by a mass of users, but it is more efficient if they can be automatically derived from users' interactions with the system. The approach proposed in this chapter focuses on generating the epistemology from an exploratory or informational search task, which is intended to find information about a topic [138]. In contrary to a navigational search task, which is intended to find specific

Web resources that the user has already had in mind (e.g., searching for books written by a specific author) and can be completed by generating a single or couple of easy-to-formulate queries, an EIS task usually requires to generate a set of queries and click to view a number of Web page URLs retrieved by each query. The main reason is that the user performing an EIS task is unlikely an expert in the topic domain and consequently it is non-trivial for them to formulate a query to precisely describe the task goal. As such, they have to iterate between generating queries and view the retrieved results until they have acquired the information that matches the task goal, i.e., results relevant to the topic, through the trial-and-error search process.

If a query does not precisely describe the goal of an EIS task, retrieving the results only matching the query terms and ranking them simply according to the term frequency cannot achieve the task goal. For users conducting EIS tasks, it is imperative that a query can retrieve Web pages that are relevant to the search goal, regardless of term match, and that are ranked according to their relevance to the search goal, regardless of term frequency. Moreover, the coherent queries and the ranked Web pages that are relevant to the search goal should be contained in the epistemologies generated from these search tasks.

One of the major technologies to addressing this issue is query expansion [135], which reformulates a seed query through techniques such as term re-weighting [136], association rules [177], or ontologies [20]. Recent research has been focusing on the query clustering approach through analyzing the click-through information in query logs. The approach can be illustrated by a bipartite graph [16] in Figure 4.1, where vertices in

the left set $\{q_1, q_2, q_3, q_4\}$ represent queries composed of terms, e.g., query q_1 is composed of a set of terms $\{t_{11}, t_{12}, ..., t_{1m}\}$, vertices in the right set $\{u_1, u_2, u_3, u_4\}$ represent Web page URLs, and a solid edge between q_x to u_y represents u_y has been clicked by the user who issued q_x, while a dashed edge represents u_y is relevant to q_x according to the search goal but it is not retrievable by q_x through matching terms.

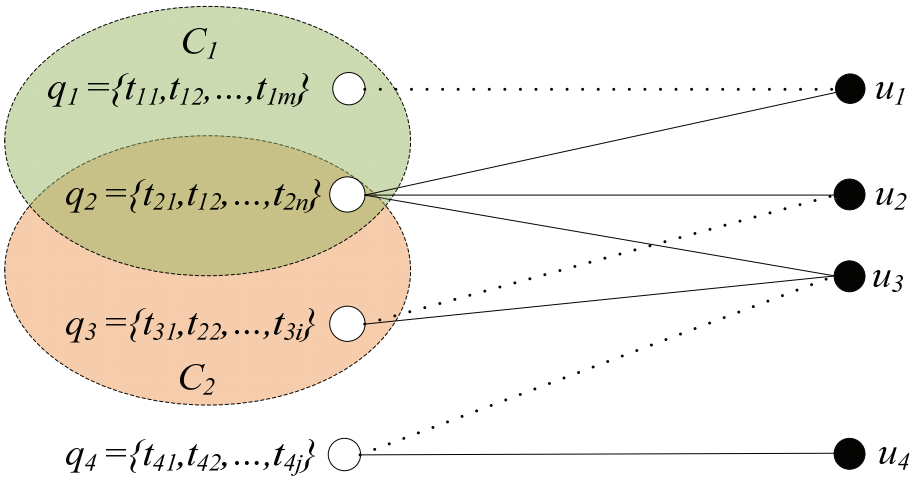

Figure 4.1 Bipartite graph representation of click-through data

There are two main query clustering techniques. One is based on the similarity between queries measured by Levenshtein edit distance [158][181], as queries with common terms are likely to be related. In Figure 4.1, queries q_1 and q_2 are placed in the same cluster C_1 using this technique because both queries contain a common term t_{12}. The other is based on the relationships between queries and Web pages [12][16], as queries returning common clicked URLs or similar Web page contents are likely to be related. In Figure 4.1, queries q_2 and q_3 are placed in the same cluster C_2 using this technique because both queries return a common clicked URL u_3.

However, the query clustering approach is unable to cluster queries that are intrinsically related but neither contain common terms nor return common clicked URLs. In Figure

4.1, suppose q_4 = "how to speed up windows xp", q_1 = "anti virus update", q_2 = "security system virus", and q_3 = "download service pack 2". Neither will q_4 be clustered into C_1 nor into C_2 because it does not contain any term in common with those used by the queries in C_1 or return any clicked URL in common with those clicked by the users who generated the queries in C_2, although q_4 is intrinsically related to both q_1 and q_3 because in reality many users who generated q_4 also generated q_1 or q_3.

More importantly, the query clustering approach does not address the issue of ranking retrieved Web page URLs according to their relevance to the goal of an EIS task. URLs are usually ranked according to the "popularity" – the number of clicks in a cluster [9]. Although some ranking techniques are available, they are independent of the query clustering approach and generally do not rank results according to the task goal.

In this chapter, we present a new probability-based approach for epistemology generation that is different from query clustering. The cornerstone of the new approach is a novel probabilistic topic model to discover the latent semantic relationships between the queries and the URLs retrieved by the queries. The topic model is based on the statistics of co-occurrence of query terms and URLs, and a variety of topics are derived to measure the similarity among the queries based on their probability distributions over these topics. The epistemology can be generated from this model as it models a user's EIS process, including query formation, results evaluation and outcome annotation. Therefore it contains all useful information that is relevant to the use's search goal, such as the semantically related queries, the URLs retrieved by the queries and the annotation associated with the URLs added by the user.

The novelty of this approach is twofold. One is that it can discover related queries that cannot be clustered by the query clustering approach, because it measures the relevance of queries at the term level (i.e., the probability of terms being associated with the same topic), while the query clustering approach does that at the query level (i.e., comparing queries in their entireties in order to find out common terms or URLs). In our approach, the queries are related if they have similar probability distributions determined by the probability distribution of each term in the queries. The other novelty of the approach is that it can rank the retrieved Web page URLs according to their relevance to the goal of an EIS task based on the similarity of the probability distributions among the URLs and the queries over all latent topics. Further, we model the process of EIS with social tags (e.g., social bookmarks) to infer users' actual goals so that more relevant topics could be derived.

We have conducted a set of experiments to validate this approach and the results have shown that our approach can significantly improve the performance of an EIS task in terms of search accuracy and search efficiency defined in Section 4.4.

4.2 Related Work

Query reformulation has been studied extensively in the IR community. One of the earliest notions is Rocchio's classic query reformulating scheme [136], where query terms were re-weighted based on feedback relevance. Recent research efforts include local context analysis based on pseudo-relevant documents [187] (e.g., top-ranked documents), mining term association rules for automatic global query expansion based on

selected corpus [177] (e.g., TREC collections), and ontology-based query expansion using external knowledge resources [20] (e.g., WordNet). These efforts share one thing in common: they do not exploit the real search activities performed by real users, e.g., the queries they generated and the URLs they clicked.

User interactions recorded in user logs have been used to improve the performance of query expansion [39]. In particular, recent research has been focusing on clustering queries based on the click-through information in query logs in order to discover the relevance of the queries frequently submitted to an SE. The objective is to expand a new query with related existing queries that were generated by other users. In Smyth's work [158], queries were treated as sets of unique terms and their similarities were computed using the Jaccard measure. In Beeferman's work [16], a bipartite graph was first constructed from the click-through data to represent the queries and the retrieved documents and a graph-based agglomerative iterative clustering method was then used to merge vertices of the graph continually until a termination condition reached. In Baeza-Yates' work [12], a weighted graph derived from the query-click bipartite graph was used to infer the semantic relations among the queries in a vector space. Wen's work [181] used a density-based method to estimate the similarity between queries by combining query content and click-through information. Although these research efforts have exploited the real search activities performed by real users, some intrinsically related queries still could not be clustered together because they analyze query relevance at the query level and queries in their entireties are generally sparse. In contrast, our approach analyzes query relevance at the term level and therefore it can discover more related queries because terms are less sparse than queries.

Moreover, they do not rank the Web page URLs retrieved by the queries in a cluster according to their relevance to the general search goal of that query cluster. Instead, queries in the same cluster are treated equally and each query's results are ranked individually. Optimizing or re-ranking search results from multiple queries by analyzing click-through data is a separate research endeavor independent of query clustering [1][77][130][173][184][189]. A representative example of such endeavor is "learning to rank" with Neural Network or Support Vector Machine. For example, in Joachims's work [77], a ranking function could learn from the implicit feedback in a SE's click-through data to provide personalized ranking of search results. In contrast, ranking of search results from multiple queries according to their relevance to the general goal of an EIS task is built into our approach using the same probabilistic topic model.

Our topic model is extended from LDA (Latent Dirichlet Allocation) [26] - a probability-based language model that can be used to find the latent semantic relationships between the words in a document. LDA and its extensions and variants have emerged as a useful family of graphical models with many interesting applications mainly for natural language processing [24][25]. Recently, there are some attempts to use LDA for IR. For example, Hörster et al. [70] proposed an LDA model that can be used to project an image onto a latent space, where image similarities are measured for the purpose of retrieving related images. Wei and Croft [178] incorporated LDA into a language modeling IR framework in order to improve the performance of ad-hoc IR.

Our topic model RTU-LDA is a significant extension to the original LDA in two folds. First, it allows a document to be associated with any number of labels, while LDA does

not support labeled documents at all and sLDA [25], an LDA extension, limits a document to be associated with only a single label. Second, it supports structured documents, while the original LDA only supports documents in plain text. In particular, our document model has a hierarchical structure for recording an EIS process: a document comprises a number of queries; each query is associated with a number of clicked URLs; and each URL is labeled by a number of tags. The technical motivations are described in more detail in Appendix D.

4.3 A Probabilistic Topic Model for Epistemology Generation

4.3.1 The Topic Model with Social Tags

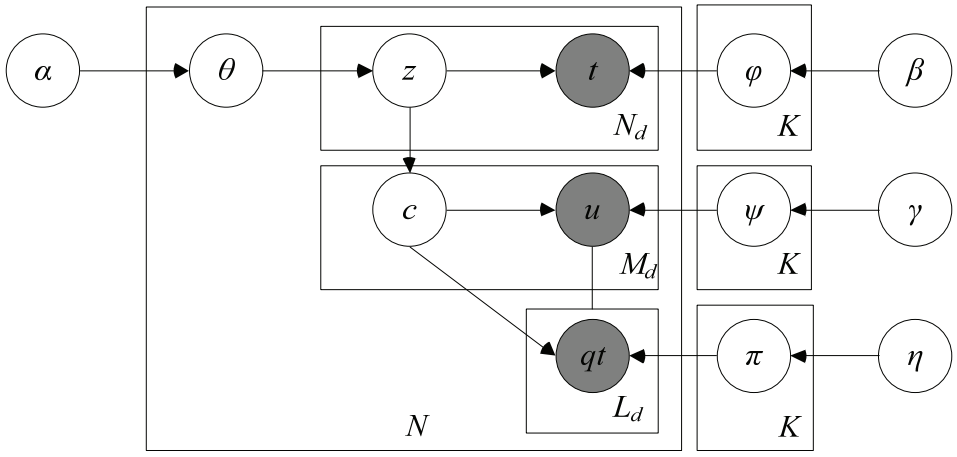

Figure 4.2 RTU-LDA's graphical representation

As a significant extension of LDA, our topic model RTU-LDA (Relevance for Terms and URLs) allows a document (the epistemology) to be hierarchically structured and tagged with any number of labels. Each epistemology records a user's EIS activity: first choosing terms to formulate a series of related queries, then clicking some of the

retrieved URLs to view the Web pages retrieved by these queries, and finally adding some tags to annotate some of the viewed Web pages.

The graphical representation of the RTU-LDA model is shown in Figure 4.2, while the notations are given in Table 4.1:

Table 4.1 Notations used in RTU-LDA

Symbol	Description
N	number of documents; each document is composed of the click-through information and the tags for the clicked URLs collected from an EIS process
K	number of topics
qt	query term
u	Web page URL
t	tag
z, c	topic
θ	document's topic distribution
π	topic's query term distribution
ψ	topic's URL distribution
φ	topic's tag distribution
L_d	number of query terms in document d
M_d	number of URLs in document d
N_d	number of tags in document d
η	query term hyperprior
γ	URL hyperprior
β	tag hyperprior

In RTU-LDA, query terms, URLs, and tags are all observed variables, while the hidden variable – the latent topic - can be discovered by the observation of tags in the training process. The generative process of a document (the epistemology) is abstracted from an EIS process: first selecting tags that can express the user's search goal, then selecting URLs labeled with these tags, and finally selecting query terms used to retrieve these URLs. URLs not labeled with any tag and query terms not associated with any clicked URL will be selected with the probabilities specified by a distribution.

Formally, RTU-LDA assumes each document d has topic proportions θ_d that are sampled from a Dirichlet distribution. For each topic z, a collection of tags t_d are selected from a topic-specific multinomial distribution φ_z. Topic c is sampled from topics $z_d = \{z_{dn}\}_{n=1}^{N_d}$ with a multinomial distribution, in which $p(c = k) = \frac{N_{kd}}{N_d}$, where N_{kd} is the number of tags that are assigned to topic k in the dth document. For each topic c, a collection of URLs u_d are sampled from a topic-specific multinomial distribution ψ_c, and a collection of query terms qt_d are sampled from a topic-specific multinomial distribution π_c.

RTU-LDA assumes the following generative process for a collection of documents, each of which has the structure of $\{(t_d, u_d, qt_d)\}_{d=1}^{D}$.

1. Draw topic proportions $\theta_d \sim \text{Dirichlet}(\alpha)$.
2. For each tag t_{dn}, $n = 1, \ldots, N_d$,
   (a) Draw topic $z_{dt} \sim \text{Multinomial}(\theta_d)$
   (b) Draw tag $t_{dn} \sim \text{Multinomial}(\varphi_{zdn})$

3. For each URL u_{dm}, $m = 1, \ldots, M_d$,

   (a) Draw topic $c_{dm} \sim \text{Multinomial}(\{\frac{N_{kd}}{N_d}\}_{k=1}^{K})$

   (b) Draw URL $u_{dm} \sim \text{Multinomial}(\psi_{c_{dm}})$

4. For each query term qt_{dl}, $l = 1, \ldots, L_d$,

   (a) Draw topic $c_{dm} \sim \text{Multinomial}(\{\frac{N_{kd}}{N_d}\}_{k=1}^{K})$

   (b) Draw query term $qt_{dl} \sim \text{Multinomial}(\mu_{c_{dm}})$

Based on the RTU-LDA model, the likelihood of all query terms $QT = \{qt_d\}_{d=1}^{D}$ being associated with specific topics is the joint distribution of all variables (observed and hidden):

$$p(Z,T,C,U,QT \mid \alpha,\beta,\gamma,\eta) =$$
$$p(Z \mid \alpha)\,p(T \mid Z,\beta)\,p(C \mid Z)\,p(U \mid C,\gamma)\,p(QT \mid C,U,\eta)$$

where $Z = \{z_d\}_{d=1}^{D}$, $T = \{t_d\}_{d=1}^{D}$, $C = \{c_d\}_{d=1}^{D}$, and $U = \{u_d\}_{d=1}^{D}$.

4.3.2 Estimation of Parameters

Inference of all latent topics Z from existing documents entails learning the model parameters θ, φ, ψ, π - the distributions over topics, tags, URLs, and query terms - respectively.

Although exact computation of these parameters is intractable, several approximation methods have been proposed to solve similar parameter estimation problems. We adopt

Gibbs Sampling [56] - a special case of Markov-chain Monte Carlo methods that estimate a posterior distribution of a high-dimensional probability distribution - to solve the parameter estimation problem in the RTU-LDA model. The sampler draws from a joint distribution $p(x_1, x_2, x_3, ..., x_n)$ assuming the conditionals $p(x_i \mid x_{\neg i})$ are known, where $x_{\neg i} = (x_1,...,x_{i-1},x_{i+1},...,x_n)$.

We derive the Gibbs sampler's update equation for the hidden variables from the joint distribution and arrive at:

$$p(z_i = j \mid z_{\neg i}, T, C, U, QT) \propto$$
$$\frac{N_{\neg i,j}^{(t_i)} + \beta}{\sum_{t_i=1}^{N_d} N_{\neg i,j}^{(t_i)} + T\beta} \frac{N_{\neg i,j}^{(d_i)} + \alpha}{\sum_{k=1}^{K} N_{\neg i,k}^{(d_i)} + K\alpha}$$
$$\frac{N_{\neg i,j}^{(u_i)} + \gamma}{\sum_{u_i=1}^{M_d} N_{\neg i,j}^{(u_i)} + U\gamma} \prod_{u_i=1}^{M_d} \frac{M_{\neg i,j}^{(u_i)}}{M_t^{(u_i)}} \frac{N_{\neg i,j}^{(d_i)}}{N_d}$$
$$\frac{N_{\neg i,j}^{(qt_i)} + \eta}{\sum_{qt_i=1}^{L_d} N_{\neg i,j}^{(qt_i)} + QT\eta} \prod_{qt_i=1}^{L_d} \frac{L_{\neg i,j}^{(qt_i)}}{L_u^{(qt_i)}} \frac{N_{\neg i,j}^{(d_i)}}{N_d}$$

where $N_{\neg i,(\cdot)}^{(\cdot)}$ is a number excluding the current position assignments of z_i, e.g., $N_{\neg i,j}^{(t_i)}$ is the number of tag t_i generated by the *j*th topic excluding the current position, $M_t^{(u_i)}$ is the total number of tags associated with URL u_i, and $M_{\neg i,j}^{(u_i)}$ is the number of tags associated with URL u_i that are assigned to topic z_j excluding the current position. Similarly, $L_u^{(qt_i)}$ is the total number of URLs that can be retrieved by a query including query term qt_i and $L_{\neg i,j}^{(qt_i)}$ is the number of corresponding URLs that are assigned to topic z_j excluding the current position.

For any simple sample, we can estimate θ, φ, ψ, π using:

$$\theta_j^{(d_i)} = \frac{N_j^{(d_i)} + \alpha}{\sum_{k=1}^{K} N_k^{(d_i)} + K\alpha}, \ \varphi_{j,t_i} = \frac{N_j^{(t_i)} + \beta}{\sum_{t_i=1}^{N_d} N_j^{(t_i)} + T\beta}$$

$$\psi_{j,u_i} = \frac{N_j^{(u_i)} + \gamma}{\sum_{u_i=1}^{M_d} N_j^{(u_i)} + U\gamma}, \ \pi_{j,qt_i} = \frac{N_j^{(qt_i)} + \eta}{\sum_{qt_i=1}^{L_d} N_j^{(qt_i)} + QT\eta}$$

4.3.3 Retrieval and Ranking of URLs

Based on the RTU-LDA model, we can compute the probability of each query being associated with each of the topics according to the probability of each term of the query being associated with the topic. Given a query q composed of N terms, i.e., $q = \{qt_1, qt_2, \ldots, qt_n\}$, the probability of q being associated with all K topics $T = \{T_1, T_2, \ldots, T_k\}$ can be expressed as a $N \times K$ matrix:

$$p(q \mid T) = \begin{bmatrix} p(qt_1 \mid T_1) & p(qt_1 \mid T_2) & \ldots & p(qt_1 \mid T_K) \\ p(qt_2 \mid T_1) & p(qt_2 \mid T_2) & \ldots & p(qt_1 \mid T_K) \\ \ldots & \ldots & \ldots & \ldots \\ p(qt_n \mid T_1) & p(qt_n \mid T_2) & \ldots & p(qt_n \mid T_K) \end{bmatrix}$$

We then get the probability distribution of q over all K topics using the Bayes law:

$$\begin{bmatrix} p(T_1 \mid q) & p(T_2 \mid q) & \ldots & p(T_K \mid q) \end{bmatrix} \propto$$
$$\begin{bmatrix} \frac{\sum_{i=1}^{n} p(qt_i \mid T_1)}{\sum_{k=1}^{K}\sum_{i=1}^{n} p(qt_i \mid T_k)} & \frac{\sum_{i=1}^{n} p(qt_i \mid T_2)}{\sum_{k=1}^{K}\sum_{i=1}^{n} p(qt_i \mid T_k)} & \ldots & \frac{\sum_{i=1}^{n} p(qt_i \mid T_K)}{\sum_{k=1}^{K}\sum_{i=1}^{n} p(qt_i \mid T_k)} \end{bmatrix}$$

Another query $q' = \{qt'_1, qt'_2, \ldots, qt'_m\}$ is related to q if they have similar probability distributions over the topics, which can be measured using the Kullback-Leibler Divergence:

$$D(q' \| q) = \sum_{k=1}^{K} p(T_k \mid q') \log \frac{p(T_k \mid q')}{p(T_k \mid q)} \qquad (1)$$

Therefore we can expand an imprecise query with related existing queries through discovering the relevance between two queries based on their KL-divergence computed by formula (1). Two queries are related if their divergence value is small, i.e., if their query terms have similar probabilities of being associated with the same topic.

The probability of URL u being associated with all the K topics is:

$$p(u \mid T) = \left[p(u \mid T_1) \quad p(u \mid T_2) \quad \ldots \quad p(u \mid T_k) \right]$$

We also get the probability distribution of u over all K topics:

$$\left[p(T_1 \mid u) \quad p(T_2 \mid u) \quad \ldots \quad p(T_K \mid u) \right] = \left[\frac{p(u \mid T_1)}{\sum_{k=1}^{K} p(u \mid T_k)} \quad \frac{p(u \mid T_2)}{\sum_{k=1}^{K} p(u \mid T_k)} \quad \ldots \quad \frac{p(u \mid T_K)}{\sum_{k=1}^{K} p(u \mid T_k)} \right]$$

The relevance between q and u is measured by their similarity of probability distributions over the topics:

$$D(u \| q) = \sum_{k=1}^{K} p(T_k \mid u) \log \frac{p(T_k \mid u)}{p(T_k \mid q)} \qquad (2)$$

If q's terms and u have similar probabilities of being associated with all topics, URL u is likely to be relevant to query q.

We then use the weighted Borda count method [8] as a rank aggregation method to combine the URL list retrieved through the RTU-LDA model with the lists retrieved by a conventional SE. A score is computed and assigned to each candidate URL according to the URL's position in each ranked lists and all candidate URLs are then ranked according to their total scores.

In summary, the epistemology generation using the RTU-LDA model is through the following five steps:

1. When a user generates a new query, it will be expanded with related existing queries discovered using formula (1). These queries are those whose terms have high probabilities of being associated with the topic that the terms of the new query are about.
2. The URLs associated with these queries will be discovered using formula (2).
3. Ranking scores will be computed and assigned to each of the URLs retrieved at step 2) as well as by a conventional search engine.
4. All URLs are ranked according to their scores. Top ranked URLs are those that have the highest probabilities of co-occurrence in the topic that the terms of the new query are about.
5. The related queries, the retrieved and ranked URLs, and the tags associated with these URLS are hierarchically organized (e.g., in XML format) together to generate an epistemology, which can record this exploratory search process and facilitate future retrieval.

Figure 4.3 The epistemology generation algorithm

4.4 Experiments

The effectiveness of our approach for epistemology generation is evaluated by a set of experiments validating that: 1) the proposed query expansion approach based on the RTU-LDA topic model can retrieve more relevant results than alternative approaches, and 2) the retrieved results can be ranked according to their relevance to the goals of users' EIS tasks.

Our experimental data is derived from two real-world datasets: query logs from a commercial SE AOL and a social annotation dataset from Delicious. The experiments compare the probability-based approach with the following alternative approaches:

1. *Baseline approach*: the approach used by the AOL SE to retrieve and rank results. It is worth clarifying that because AOL SE's retrieval and ranking techniques are not public, the baseline approach is actually derived from AOL query logs.
2. *Query-clustering approach*: this approach uses a query-clustering algorithm similar to the one proposed in [9] to cluster semantically connected queries. The degree of similarity between two queries is decided by the fraction of common terms in the clicked URLs. The URLs returned by all the queries in a cluster are ordered according to the number of clicks occurring to them, which leads to the popularity ranking of URLs. The new ranking of URLs is boosted by combing the popularity ranking and the original ranking returned by the SE.
3. *Learning-to-rank approach*: for a sequence of queries generated by a user, this approach uses the algorithm developed in [130] to generate preference judgments

about the relative relevance of the documents retrieved by an individual query and also those retrieved by different queries in the sequence. The ranked retrieval function is learned from the preference judgments using a ranking SVM.

4.4.1 Datasets

The AOL query logs used in the experiments contain about 20 million search queries from 657,426 users [123]. For each query, the URLs clicked by the user and the ranking of each clicked URL on the result list were recorded in the logs. We first removed the stop words and stemmed the query terms using the Porter stemming algorithm [128]. For the purpose of preserving privacy and removing noises in the training process, we only kept those queries that contain no rare terms (i.e., terms occurred less than five times).

Because our objective is to compare the performance of the proposed probability-based approach with alternative ones in terms of supporting EIS tasks, we randomly chose 100 users from the dataset and manually extracted two EIS tasks accomplished by each user as the test dataset, i.e., 200 EIS tasks in total for the experiments. Each task is composed of a set of inter-related queries $\{q_1, q_2, \ldots, q_n\}$ (n > 1) and a list of URLs $[u_1, u_2, \ldots, u_m]$ (m > 1) retrieved by the set of queries and ranked according to their relevance to the goal of the user's EIS task (the relevance judgments were provided by the participants of these experiments). Well-defined metrics are used to evaluate the performance of each approach in accomplishing these tasks.

The remainder of the query logs were used for training the model and the algorithms. For the probability-based approach, a user's query terms and clicked URLs are treated as a

document, which might be labeled with social annotations. The social annotation dataset used in the experiments is DeliciousT140 from Zubiaga [207], which was created with data retrieved from the social bookmarking site Delicious and various Web sites. This dataset contains 144,574 unique URLs and 67,104 different corresponding social tags retrieved from Delicious on June 2008. For a document containing a set of query terms and URLs, if a URL could be retrieved from this social annotation dataset, all its associated tags would be used as the labels of the document for training.

4.4.2 Methodology

The experiments were conducted to evaluate the performance of the proposed approach and alternative approaches in terms of the model quality and the retrieval quality.

For the model quality, we compared the proposed RTU-LDA model with the existing LDA and sLDA [25] models, in terms of the ability to identify latent topics for discovering the relevance between URLs and query terms. The LDA model is based on the co-occurrence probabilities of the query terms and URLs. The click-through data is treated as a bag-of-words document, where "words" are query terms or URLs. The sLDA model selects a single tag for each document as the label for supervised training. The selected tag is the most frequently used one among all the tags for annotating the document.

For the retrieval quality, we compared the retrieval results of the probability-based approach with those of the baseline, query-clustering, and learning-to-rank approaches, in terms of the ability to retrieve and rank URLs that are relevant to a search goal.

Model quality:

To evaluate the quality of a model, we measure its performance in discovering the relevance between the URLs and a query's terms using the perplexity of held-out query-related URLs for the query. Perplexity is widely used in language modeling and can be defined as follows [202].

$$Perplexity(D) = \exp\{-\frac{\sum_{q \in D} \sum_{u \in N_q} \ln P(u \mid q)}{\sum_{q \in D} N_q}\}$$

where D is a test document, q is a query in the test document, N_q is the set of URLs that are related to q, and $P(u \mid q)$ is the probability of URL u being associated with query q inferred by the model.

As perplexity is algebraically equivalent to the inverse of the geometric mean per-word likelihood, a low perplexity represents high performance. As LDA-based models are sensitive to the number of topics, we can optimize the models by analyzing the influence of perplexity using different topic numbers.

Retrieval Quality:

A number of methodologies and metrics have been used to evaluate the quality of retrieval [11]. In these experiments, we specifically want to evaluate how the retrieved results have satisfied the goal of an EIS task in terms of search accuracy and search efficiency.

- *Search Accuracy*

We measure the search accuracy using the following independent metrics:

- Relevance accuracy: given the search results of the test dataset, we measure how many relevant documents have been retrieved. It should be pointed out that the relevance is relative to the overall goal of an EIS task rather than to each individual query used in the search process. Evaluation metrics of Precision, Recall and MAP [29] are used to measure relevance accuracy.

  Precision is the fraction of retrieved documents that are relevant, while Recall is the fraction of relevant documents that have been retrieved.

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(relevant \mid retrieved)$$

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(retrieved \mid relevant)$$

  Recall and Precision usually contradict each other: low Precision means many results are irrelevant and low Recall means many relevant results have not been retrieved. An EIS task would have a low Recall if retrieved results were only those that match some of the query terms. Our approach can retrieve more results that are relevant to a search goal, thus achieving high Recall, while keeping reasonable high Precision at the same time.

- Ranking accuracy: given the search results of the test dataset, we measure how accurately these results are ranked according to the

relevance to the search goal. Again, the relevance is relative to the overall goal of an EIS task rather than to each individual query used in the search process. NDCG [74] is used to measure ranking accuracy.

- *Search Efficiency*

  Search efficiency measures how effective the queries are. In particular, an EIS task would have high search efficiency if the user generated few queries but each query retrieved many relevant documents.

  The following SE (Search Efficiency) metric is defined to measure search efficiency:

$$SE = \frac{1}{N}\sum_{i=1}^{N} RUN_k(q_i)$$

  where N is the total number of queries generated by the user to achieve their search goal and $RUN_k(q_i)$ is the number of relevant URLs in each query's top-k retrieved results. To obtain the statistical significance in measuring the search efficiency, we chose queries from the test dataset randomly in three different orders and calculated the average SE scores.

4.4.3 Results

In this section, we will present and discuss the results of our experiments.

Model quality:

It is possible for us to choose the right number of topics for grouping query terms and Web pages through the model quality evaluation. The reason is that the performance of RTU-LDA and other LDA variants is largely dependent on the number of topics chosen. We first decide the range of the topic number, and then select the number which achieved the best results in the experiment. We hope it could reflect the breadth of the concepts available in the Internet, as the topics will be used to classify the searches of documents on the Web.

We have utilized well-known Web categories for classifying searches instead of constructing document classification from the scratch. The first category is Yahoo! Answers, where users can select a topic from the predefined taxonomy and assign one category to each of their question. The second category is the Yahoo! Directory hierarchy, where millions of unique pages had been manually classified into a hierarchy of categories by trained professional web editors. An advantage for utilizing Yahoo! Answers and Yahoo Directory is that many Internet users are familiar with them. However, the category hierarchies in Yahoo! Answer and Yahoo! Directory have too many levels and the topics at lower levels are too specific for categorizing searches. Therefore in this experiment, we utilized only the first two levels of these two categories. A simple program was written to fetch the first two levels from Yahoo! Answer and the Yahoo! Directory. Some manual adjustment of the topic numbers was made to reduce the noises. For example, there are some topics such as “Regional”, “By Region”, “Browse by Region”, in Yahoo Directory. These topics were removed as they are not considered to be very useful for classifying searches.

In total, Yahoo! Answer contains 26 top-level categories, each branching into 10.2 (±8.1) second-level categories on average. Yahoo! Directory has 13 categories in the first level, each comprising 14.3 (±7.6) subcategories in the second level on average. Therefore we expect the topic number would approximate to the numbers of overall second-level categories of these two categories, i.e., 265 and 186 respectively.

Figure 4.3 shows the perplexity tested on the held-out sample of each of the LDA-based models for different topic numbers K = 20, 50, 100, 150, 200, 250, 300. It is clear that for any fixed number of latent topics K, the RTU-LDA model has achieved the lowest perplexities, indicating that it is most capable to discover the latent relevance between URLs and query terms among the three LDA-based models.

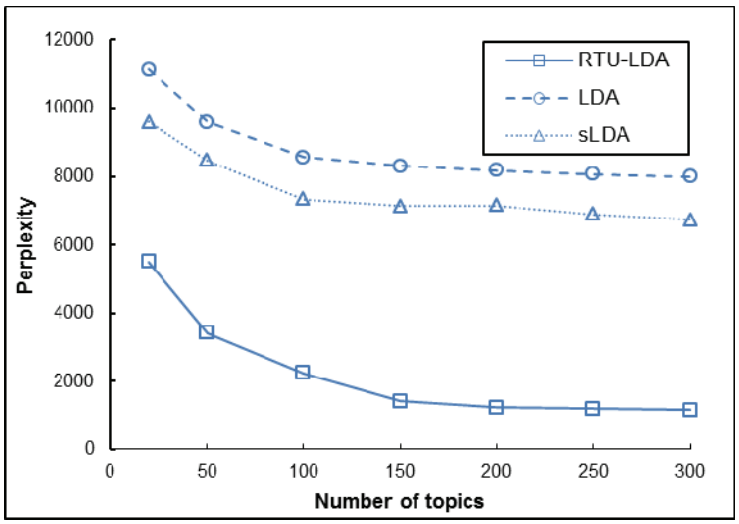

Figure 4.4 The perplexities of different models for K = 20, 50, 100, 150, …, 300

The high perplexity of LDA is probably due to the polysemy of query terms and the ambiguity of queries in EIS tasks. For example, given the top ten words associated with a single topic learned by LDA: “home”, “speed”, “monitor”, “real”, “free”, “race”, “estate”, “window”, “furniture”, and “download”, LDA will probably infer that the queries of

"home estate" and "speed up windows" are related, which are obviously not. In contrast, RTU-LDA will derive one topic where "home" and "estate" are top words and another topic where "window" and "download" are top words.

The perplexity of sLDA is slightly better than that of LDA because the document label has helped find the correct topic for query terms. For example, the two queries of "home estate" and "speed up windows" will be assigned by sLDA to two different topics because the documents associated with the two queries are tagged by two different labels. However, sLDA only uses a single label to tag each document, which may be inadequate for most documents as they may contain more than one topic. This explains why the perplexity of sLDA is a lot worse than that of RTU-LDA, which, in contrast, has the ability to tag a document with multiple labels.

Figure 4.4 also shows that the perplexity of RTU-LDA becomes steady when the number of topics reaches 200. Therefore we set K=200 as the number of topics for deriving latent topics and estimating parameters in the RTU-LDA model. This result also agrees with our supposition in the beginning of the experiment. To be consistent, we have set the same model hyper parameters (α=50/K, β=0.1) for all models in the experiments.

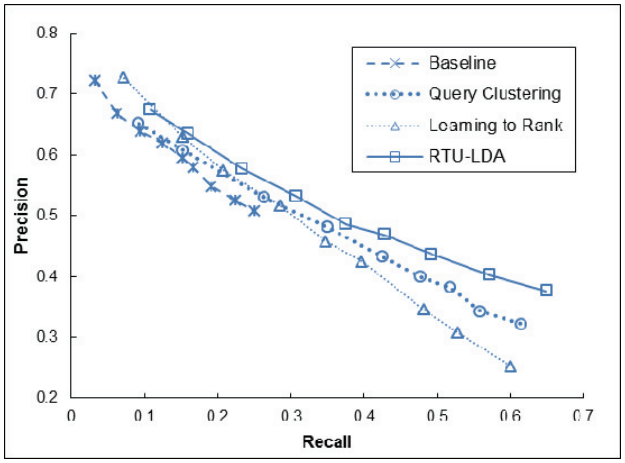

Figure 4.5 The precision-recall graph of the four approaches

Precision and Recall:

Figure 4.5 shows the precision-recall graph of the four approaches. It can be seen that RTU-LDA-based approach outperforms the baseline approach in most cases except at the low recall level. Because our objective is to retrieve more documents that are relevant to the goal of an EIS task, we expect high precision at the high recall level. The baseline approach is not able to achieve high recalls due to the fact that the SE has only retrieved the documents that match some of the query terms.

The query clustering approach has improved the recall through retrieving more documents that are relevant to the search goal. However, since it only clustered queries with common terms in the clicked URLs, documents with latent relevance to the queries could not be retrieved. Therefore this approach's precisions are worse than RTU-LDA's, either at low or high recall levels. The learning to rank approach is better than the baseline because it can include documents originally not present in the initial search results through learning. However, it is not easy to infer preference judgments from query logs. Compared to the RTU-LDA-based approach, this approach's performance is worse

especially at high recall levels, probably because the ranking SVM is more dependent on the proper preference judgments and the training data, while RTU-LDA is a generative probabilistic model and can deal with unseen queries more accurately.

Table 4.2 The MAP scores of the four approaches

Approaches	MAP	%change by Baseline
Baseline	0.2204	-
Query Clustering	0.2536	+15.6%
Learning to Rank	0. 2627	+19.2%
RTU-LDA	0. 2779	+26.1%

Table 4.2 shows the MAP scores of the four approaches, which further illuminate that the RTU-LDA-based approach has significantly improved the overall precision of EIS tasks, as compared to the query clustering and learn to rank approaches.

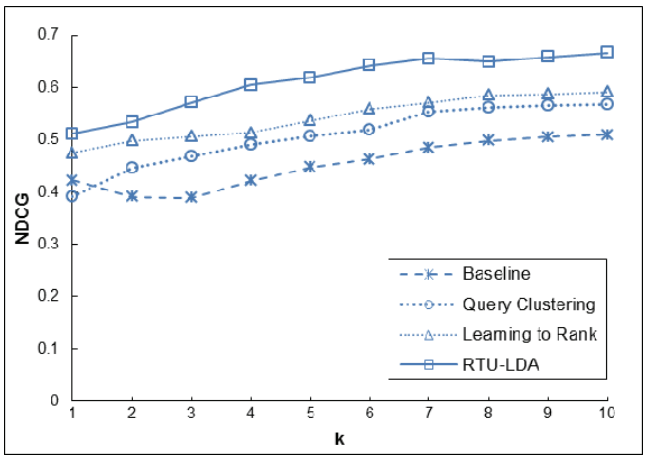

Figure 4.6 NDCG@k of the four approaches

Ranking accuracy:

Figure 4.6 shows the NDCG@k (k=1, …, 10) of the four approaches. It is clear that RTU-LDA has achieved the best ranking accuracy in terms of the relevance of the retrieved URLs to the search goals, while the baseline is the worst and the query clustering and learning to rank approaches generally outperform the baseline. RTU-LDA uses the discovered latent semantic relationships between URLs and query terms to identify the URLs that are most relevant to the search goal. Ranking in the query clustering approach is only based on URL popularity, which cannot accurately reflect the relevance between URLs and query terms. The learning to rank approach, albeit using the rank SVM to learn ranked retrieval functions, mainly evaluates the ranking functions based on click counts. Therefore both approaches tend to rank frequently clicked URLs higher, while RTU-LDA ranks the results based on the inferred latent relevance between URLs and query terms, thus achieving high ranking accuracy.

Search efficiency:

Table 4.3 shows the SE values of the four approaches. It can be seen that RTU-LDA has achieved a significant 36.8% improvement over the baseline, while the query clustering and learning to rank approaches have achieved 25.7% and 28.1% improvements over the baseline respectively. Generally speaking, compared to a baseline SE, the RTU-LDA-based SE requires significantly less number of queries - with each query retrieving significantly more relevant results - to achieve the goal of an EIS task. Therefore, the user can save a lot of time reformulating imprecise queries, clicking URLs and viewing Web pages that are irrelevant to their search goal at all. Moreover, the epistemologies

generated based on RTU-LDA contain the queries and URLs closely related to the search goal, which will facilitate the search processes of many other users.

Table 4.3 The SE (Search Efficiency) scores of the four approaches

Approaches	SE	%change by Baseline
Baseline	2.53	-
Query Clustering	3.18	+25.7%
Learning to Rank	3.24	+28.1%
RTU-LDA	3.46	+36.8%

4.5 Summary

In this chapter, we discussed the issues related to epistemology generation in the proposed social search framework to supporting EIS tasks. The generated epistemology should contain queries, URLs and tags that are relevant to the search goal. Query clustering approach can cluster queries based on their lexical similarity or common clicked Web page URLs. But many intrinsically related queries could not be clustered due to the sparse nature of the query space. Moreover, the retrieved Web pages are not ranked according to their relevance to the goal of an EIS task.

The proposed approach is based on a novel RTU-LDA topic model to discover the latent semantic relationships between query terms and Web pages. The epistemology correlated with an EIS task is generated through modeling the user's search process. It is because that in this approach, not only can related queries be discovered based on the similarity of

probability distributions over topics, but also the retrieved Web pages can be ranked based on the similarity of probability distributions of the Web pages and the queries.

The performance of the proposed approach has been evaluated by conducting experiments on real-world datasets. The results have shown that the RTU-LDA model is able to accurately discover the latent semantic relationships between queries and Web pages. The RTU-LDA-based approach can thereby outperform alternative ones such as the baseline, query clustering approach, and learning to rank approaches in terms of both the search accuracy and the search efficiency.

In future research directions, it would be interesting to investigate a further improvement of ranking accuracy by harnessing sentiment tags [192] as the quality of Web pages can be indicated by these tags. Our topic model and the training algorithms could be further optimized as the time and space complexities are crucial for online Web search.

Chapter 5

Epistemology Search

In most EIS tasks, formulated queries may often fail to precisely capture users' real intents, and consequently retrieved results may not adequately satisfy their information needs. Existing query suggestion and traditional IR techniques mainly revolve around improving the relevance of suggested queries and result documents for a given query. They are not suitable for epistemology retrieval because information seekers usually have diverse but related information needs in their EIS tasks and epistemologies contributed by different seekers are likely to contain different knowledge sets. This chapter presents a novel social-interest-directed technique to addressing the critical epistemology retrieval issue in the proposed social search solution. The experimental results show that this new technique outperforms other methods for epistemology retrieval in that both the suggested queries and retrieved epistemologies are diversified to satisfy users with different information needs, while they are still semantically relevant to the given query.

5.1 Introduction

A critical issue in the social epistemology solution is epistemology retrieval, that is, how to retrieve social epistemologies that are relevant to the goal of an EIS task. Since it is non-trivial to precisely formulate a query for such a task, there are two ways to improve the effectiveness of epistemology retrieval: first, query suggestion can be used to suggest

(likely more precise) queries that were used by other seekers and are related to an imprecise query generated by the current seeker; second, the search results can be re-ranked so that the top-ranked epistemologies would more likely be able to meet the user's information need.

On the one hand, as prosumers have contributed their interactions with SEs in our epistemology repository, we could gain insights into collected user activities and thereupon present with the current seeker relevant and high-quality queries used by others. On the other hand, as prosumers have diverse but related information needs in their EIS tasks and epistemology contributed by different prosumers are likely to contain different knowledge sets, both suggested queries and top-ranked epistemologies should also be diverse so that some of them are close to a specific information need.

Unlike classic IR, where a user's information need is usually clear and straightforward; in an EIS task, it is difficult for a user to formulate a query that precisely describes their information need, because the information need is vague in the beginning and gradually clarified by iteratively submitting queries and examining the retrieved results. For each informational query in such a process, there is inherent ambiguity in it and the best way to address the ambiguity is to allow for diversity of suggested queries and search results so that the user who generated the query could pick up at lease one that really matches their information need immediately.

As traditional IR techniques mainly focus on improving the relevance of suggested queries and result documents for a given query, previous studies on query suggestion have generally overlooked the need of diversifying query suggestions for the sake of

satisfying users' diverse information needs. Most effective methods for the query suggestion technique, such as modeling the query formulation process as a Markov chain based on the query-URL click-through data, could produce a probabilistic ranking of suggestions for a given query. Consequently the queries submitted by the majority of users such as "Amazon coupons" and "Amazon promotional code" would always get higher rank than queries submitted by the minority of users such as "Amazon Elastic Computer Cloud".

The major limitation of such methods is that they only focus on the query-URL relations and cannot take advantage of the social knowledge hidden in the epistemology repository or query logs, i.e., users' different information needs and knowledge sets. Actually, from the epistemology repository or query logs not only can we know that the user typing the query "Amazon" has the highest probability of submitting the query "Amazon coupons", but also we can know that the user submitting the query "Amazon coupons" is likely to submit the query "Amazon promotional code" too; yet is unlikely to submit the query "Amazon Elastic Computer Cloud" that will probably be submitted by the user who submits the query "Amazon Web Services". Therefore, if we could suggest a list of heterogeneous queries like "Amazon coupons", Amazon Cloud Computing", and "Amazon rainforest", it will be more effective than a list of homogenous queries such has "Amazon coupons", "Amazon promotional code", and "Amazon coupon code", because it can satisfy not only the users searching for online shopping, but also the users searching for cloud computing or geographical information.

Previous work on result diversity mainly took the content-level approaches, or the implicit approaches [145] because they account for the aspects covered by the content of each individual document implicitly. For example, the retrieved documents were either re-ranked to reduce redundancy by directly comparing their content against each other or selected for improving diversity by covering the largest number of important words in their content.

Recent work on result diversity mainly takes the concept-level, or the explicit approaches [145] because they explicitly model each query's multiple concepts to diversify the search results for the sake of covering more user intents. User information needs are partitioned in concepts (a.k.a. facets, categories, and nuggets), and the search results are diversified based on the intersection of concepts contained in the queries and documents, or on a taxonomy of concepts and their distribution in relation to a particular query.

Actually, the need for diversity in results stems from diverse user intents formulated as a query with multiple concepts. Therefore the concept-level approaches can better deliver diversified search results matching users' real intents than the content-level approaches do, which diversify search results according to the content of individual documents without taking into account the queries and the user intents.

In this chapter, we contribute a new technique that gains two ends at once: achieving diversified query suggestion and concept-level result diversification concurrently. It can identify users' different intents and query concepts by mining the social interest and semantic relevance hidden in the epistemology repository or Web search logs. For instance, we can learn from the search logs that the URLs clicked by the users who

generated queries “Amazon” and “Amazon coupons” are very different from those clicked by the users who generated queries “Amazon” and “Amazon Elastic Computer Cloud”. Therefore, for an ambiguous query “Amazon”, we could suggest diversified queries in a long-tail distribution, and diversify the search results by selecting representatives from the queries and results identified with different user interest respectively. That is, the suggested queries and the selected epistemologies shown in the first page contain information not only about online shopping, but also about cloud computing or geographical information.

To discover the social interest from the epistemology repository, we perform kernel principal component analysis (KPCA) on those related queries and results generated and clicked by real users. A random walk with restart (RWR) algorithm on the query-URL bipartite graphs using both query-URL relevance and social interest is further proposed, which satisfies both the relevancy and diversity in the ranking of suggested queries and search results. A set of experiments have been conducted on real data sets and the results show that this new technique outperforms other query suggestion methods and results diversification techniques for epistemology retrieval, in that both the suggested queries and retrieved epistemologies are diversified to satisfy users with different information needs, while they are still semantically relevant to the given query.

5.2 Related work

We divide related work into those about query suggestion and those about result diversification.

5.2.1 Query Suggestion

While query suggestion has been a common feature of major commercial SEs recently, a lot of research efforts are still going on in order to suggest good queries that are closely relevant to users' information needs.

5.2.1.1 Query Similarity

Given a query string, a straightforward method to query suggestion is to compute and recommend similar queries. The similarity between queries can be measured not merely by lexical distance, since semantic analysis via a thesaurus has been proved helpful [96]. Most of these methods obtain similar queries by replacing the queries as a whole, or by substituting constituent phrases. Jones et al. [78] introduced the notion of query substitution, where similar queries and phrases are derived from user search sessions. Fonesca et al. [45] presented an approach to extracting query transactions based on query sessions and to employing association rules for query suggestion. Query semantic similarity is also estimated using information external to the query, such as mining different anchor texts pointing to the same pages for query refinement [86]. Similarly, Glance [50] measured the similarity using the overlap between the sets of retrieved documents for queries. Sahami and Heilman [140] proposed a kernel function based on the terms appearing in documents retrieved for a query to compute the similarity between queries. Temporal query patterns have also been investigated and a representative work was done by Vlachos et al. [170], who operated on the assumption that semantically related queries have similar temporal behavior of query occurrences.

5.2.1.2 Query Log Analysis

Some recent work has employed the analysis of SE log data for query suggestion. Compared to the above methods, query log analysis utilizes the rich user behavior information (i.e., queries and post-query clicking behavior) far beyond the search result content, and leverages the massive log data for building the statistical model of user-system interaction. The approach proposed by Baeza-Yates et al. [10] is to obtain a term-weight vector representation of queries from the aggregation of the URLs clicked after the query, and suggests queries with the highest similarity and attractiveness in the clusters which are clustered based on the term-weight vector representation. Another attempt of extracting information from the query was made by Zhang and Nasraoui [200], who used a complete graph to represent each user session. The arcs between consecutive queries in the same session were weighted by a dumping factor, and the values of arcs that joined the non-consecutive queries were multiplied to calculate the similarity values for the queries.

Further, for query suggestion based on exploiting the SE click-through data, there have been several successful applications of the query-click graph introduced by Beeferman and Berger [16], which is a bipartite graph that represents the user queries and visited search results. For example, Craswell and Szummer [38] described a Markov random walk model on the click graph to rank documents given a user's query. Fuxman et al. [48] used a similar approach and a query-concept mapping based on pre-defined taxonomy of concepts to recommend related queries in the context of sponsored search. Mei et al. [108] proposed a query suggestion approach based on a parameter-free random walk model that uses a computation of the hitting time on a click graph. Cao et al. [30]

proposed a context-aware query suggestion approach based on an offline learning step that clusters a click-through bipartite graph. Instead, Baraglia et al. [14] run an incremental algorithm for updating the Query Flow Graph to compute query recommendations. However, all the above approaches have not explicitly addressed the problem of query diversity as their major concern is the relevancy rather than the diversity of suggested queries.

It is worth pointing out that many techniques in query log analysis can be applied to query suggestion although they were not originally devised for that purpose. A lot of research work has been devoted to query clustering and query classification in query log analysis. For example, Wen et al. [180] applied a density-based method to cluster queries by combining query content with click-through information, while Li et al. [94] presented improving query intent classifiers by making use click graphs. Another group of query log analysis aims to understand user behavior, as exemplified by the work of White and Drucker [183] where user trails in web searches are examined.

5.2.2 Result Diversification

Although the need for diversity has long been identified since the early work [51] in IR, the problem of Web search results diversification has not been paid much attention until recently.

5.2.2.1 Content-level Approaches

For a query about a general topic, a straightforward method to improving results diversity is to maximize the sum of content dissimilarity of the retrieved documents. While

documents with similar content would cover similar facets, dissimilar documents could cover as many different subtopics as possible [197]. Most of these content-level methods directly compare the retrieved documents against one another, or measure the dissimilarity of a search result with respect to those above it in the ranked list. Carbonell and Goldstein [31] introduced an influential criterion called maximal marginal relevance (MMR) to reduce redundancy while maintaining query relevance in re-ranked documents.

Zhai et al. [196] presented an approach to modeling both the relevance and novelty in the MMR criterion using the language models, and proposed a method to measure the novelty of documents using the KL-divergence based on a risk minimization framework. Documents could also be selected sequentially according to the probability of the document's relevance conditioned on the assumption that the previous documents are not relevant to the query [35], as users only need some of the relevant documents rather than all of them. Similarly, Bookstein [27] used explicit user feedback on relevance after every document is selected.

More recently, Gollapudi and Sharma [53] proposed a set of natural axioms for diversification and utilized two well-studied algorithms to solve the facility dispersion problems. Portfolio theory from the field of finance has also been adopted, for example, Rafiei et al. [131] used a portfolio approach to increase relevance measured by the expected return and decrease similarity measured by the associated risk.

Apart from results diversification in IR, researchers have also recognized the importance of this issue in other areas. In the context of recommender systems, Alodhaibi et al. [3] used a randomized algorithm to address the maximum diversity problem of

recommendations for composite products or services. For product search in an online shopping scenario, Vee et al. [168] clustered results into buckets based on their diversity order and selected results from those buckets in order to retrieve balanced diverse results. In the multimedia domain, Leuken et al. [92] used dynamic clustering algorithms on image features to provide visually diverse result sets.

5.2.2.2 Concept-level Approaches

Some recent work has studied results diversification by exploring the multi-concept nature of an ambiguous query. Compared to the aforementioned methods based on the inter-result similarity, concept-level methods diversify search results at the topical level by making explicit use of knowledge about the topics both the query and the documents may refer to, rather than demoting the documents that are similar to the ranked ones. The approach proposed by Agrawal et al. [1] is to use a relevance measure that considers the categories of a query and documents retrieved by the query. A results set is diversified while its results cover all categories, weighted by their probability to occur. Another attempt to increase both novelty and diversity was made by Clarke et al. [37], who used "information nuggets" to represent a query and the documents retrieved by the query. The relevance is defined as a function of the nuggets contained in the user's need and previous search results.

Our method also explicitly studies the multi-concept nature of a user-generated query, but focusing on the analysis of user interest through social behavior mining. In contrast to the aforesaid methods relying on a predefined taxonomy (e.g., Open Directory Project) or the intersection of the retrieved documents, our method can not only utilize the rich user

behavior information (i.e., queries and post-query clicking behavior) that is far beyond the content of result documents but also build a statistical model of user-system interaction with the massive click-through data.

5.3 Social Interest Discovery

The prime objective of diversified query suggestion and results diversification is to satisfy users' different information needs embedded in their ambiguous queries. A way to discover the real user interest is to find out the conceptual relationship between the suggested queries or retrieved results from the click-through data contributed by numerous real users (either in SEs' Web logs or our social epistemology repository). In this section, we will first introduce the query-URL bipartite graph for clustering related queries in an exploratory search, with corresponding documents or URLs. Then we will give an algorithm based on random walk model for diversified query suggestion and results diversification using the social interest discovered by KPCA.

5.3.1 Query-URL Bipartite Graph Construction

Click-through data, which contains information about user-clicked URLs, has been intensively studied to improve SEs' performance and recently to achieve results classification. The click graph [38], a bipartite graph between queries and URLs in which edges connect a query with the URLs that were clicked by users, is an important technique for describing the information contained in click-through data. For instance,

the social epistemology repository may contain the following four clicked URLs corresponding to the four queries generated by a number of different users:

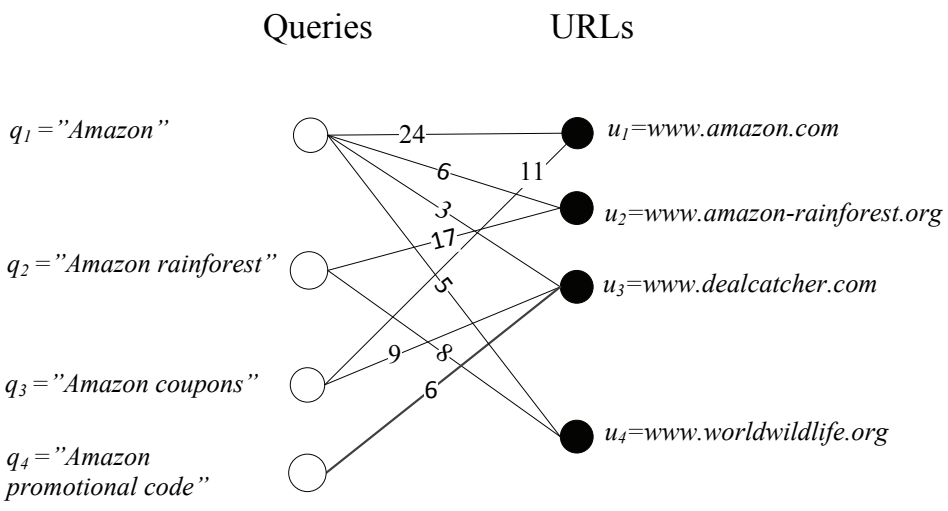

Figure 5.1 Example of a query-URL bipartite graph

Using the click graph, we can identify the conceptual relationship between a URL and an EIS task by discovering all queries associated with this URL and clustering suggestions or URLs with similar user click behavior. The rationale is that multiple queries that are co-clicked with the same URL, and URLs that are co-clicked in correspondence to the same query, should capture similar user intent. For example, URL “www.retailmenot.com” may also be clicked in correspondence to the one of the queries such as “Amazon coupons”, indicating that it is related to the URL “www.dealcatcher.com”.

Formally, let $Q = \{q_1, q_2, ..., q_N\}$ be the set of N unique queries submitted to an SE during a specific period of time. Let $D = \{d_1, d_2, ..., d_M\}$ be the set of M URLs clicked in correspondence to those queries. We construct a query-URL bipartite graph $G = (V, E)$,

where vertices V are the union of queries and URLs, i.e., $V = Q \bigcup D$, and every edge in E connects two classes of vertices: one in the query set Q and the other in the URL set D. For $q \in Q$ and $d \in D$, the pair (q, d) is an edge of E if and only if there is a user who clicked on URL d in correspondence to the query q. Each edge E is assigned a weight for click counts C_{ik}, which is the raw click frequency [38] from query q_i to URL d_k, i.e., the number of times d_k was clicked in correspondence to q_i.

Using the bipartite graph, we generate an L2-normalized feature vector F for each URL d_i in the graph. The size of the vector is the number of query vertices in the graph, and each dimension in the vector represents a query. The value for the dimension associated with query q_k is computed as:

$$\overrightarrow{f_k} = \begin{cases} \dfrac{C_{ik}}{\sqrt{\sum_j^Q C_{ij}{}^2}} & \text{if exists an edge between URL } d_i \text{ and } q_k; \\ 0 & \text{otherwise} \end{cases}$$

Then we can compute the conceptual similarity between two URLs d_i, d_j as:

$$sim(d_i, d_j) = \sum_k^Q \frac{Fi_k \cdot Fj_k}{\sqrt{\sum_k Fi_k{}^2} \cdot \sqrt{\sum_k Fj_k{}^2}} \qquad (1)$$

5.3.2 Random Walk

Using the co-click information in the bipartite graph helps find conceptually related URLs. A previous study [38] has shown that such information can also be used to derive a probabilistic model that simulates the process of formulating queries by a user for an

EIS task. We assume such a process begins with the user conceiving of their information need in the form of an imaginary document. They then formulate a query to retrieve that document. If the query's search results do not encompass documents that are close to their information need, at least one of the result documents will help them formulate another query. The process continues as they iterate between query-document and document-query transitions until at least one document meeting their information needs is finally retrieved.

The transition probabilities can be estimated from the clicks of enormous real users. Mathematically, they can be modeled as a Markov random walk between queries and URLs. Random walk starts with a vertex on the bipartite graph and then iteratively visits its neighbors with a probability proportional to the edge weights. To learn the relevance score between URLs, consider the vertices at one side, such as the URL-to-URL graph, a new random walk can be introduced by the transition probability from URL d_i to d_j:

$$p(d_j \mid d_i) = \sum_{k}^{Q} p(d_j \mid q_k) p(q_k \mid d_i) \qquad (2)$$

It is important to note that the self-transition probability exists naturally in the model.

RWR [167] is a technique that specifies a starting vertex for a walk, and at each step, the walk has a constant probability of p to jump back to its starting vertex. It has been shown that after a certain number of steps, the probability of visiting a vertex j from the starting vertex i will become stable. In other words, the user forgets their queries in the search task after some transitions.

As personalized PageRank [64] is a steady-state distribution of the random walk, we can use it to rank vertices on the graph in a query-dependent way. The corresponding linear system of personalized PageRank can be shown as:

$$R_j^{n+1} = (1-p)R_j^{(0)} + pWR_i^n \qquad (3)$$

where n is the steps of a random walk, W is the weight matrix, thus $WR_i^n = \sum_i p(v_j \mid v_i)R_i^n$.

Each entry of R_{ij} defines the relevance score of vertex v_j in relation to the starting vertex v_i. $R_j^{(0)}$ is a personalized (or query- dependent) initial values for vertex v_j. We may set $R_j^{(0)}$ to be 1 if v_j is the given query and 0 otherwise.

It can be observed that Eq.(3) can be solved iteratively by replacing R_i from the previous iteration. For example, in Figure 5.1, we can first calculate the transition probabilities from a starting query (e.g. "Amazon") to each URL normalized by the number of clicks between the query and all URLs, and the transition probabilities from all URLs to each query normalized by the number of clicks between the URL and all queries. Then the random walk will sum the probabilities of all paths of length n from the starting query to another query. Hence the query (e.g. "Amazon coupons") which has largest transition probability from the starting query will get the highest relevance score and be chosen for suggestion. Similarly, the retrieved URLs will be ranked based on the relevance score according to the transition probability from the starting query.

After convergence, the final ranking matrix $R = \{R_1, ..., R_N\}$ for N vertices is column-normalized and contains the stable relevance scores of all vertices within the graph. The only parameter in Eq.(3) is the restarting rate p which controls the behavior of the

probability distribution of neighborhood relevance scores. Higher p values have more local effect; they assign very higher scores to its nearby vertices and generally ignore the vertices far apart. Lower p values generate flatter probability distributions so that the starting vertex i is more likely to reach out to distant vertices. The parameter p is usually set to be 0.85 in previous studies [121].

5.3.3 Determine Social Interest

Because different users have different information needs, queries issued and URLs clicked by different users should be treated differently for query suggestion and the results diversification. For example, URLs clicked by a user interested in online shopping are different from URLs clicked by a user interested in cloud computing. Our solution aims to determine the different social interest according to users' "contributions" (in terms of generated queries and clicked URLs) to an entire EIS task. Then the ranking of suggestion and search results are diversified according to different users' information needs represented in the social interest.

KPCA [148] is used for social interest discovery in this work. As a non-linear extension of PCA, KPCA has proven powerful as a preprocessing step for classification algorithms. KPCA is an orthogonal basis transformation. It creates a new series of components in which the axes of the new coordinate systems point in the direction of decreasing variance, so that it can reduce the redundancy contained within the data. For example, suppose in the high school student examination, each student has scores from 12 different academic fields. These 12 fields may be correlated with two indices of intelligence associated with students, ''verbal intelligence'' and ''mathematical intelligence''. Then

these indices can be regarded as the principal components of the students. In our approach, each new component represents one aspect of the interest in an EIS task.

Mathematically, consider an exploratory search consisting of an array of *m* related queries and URLs generated and clicked by all users. A user's contribution to the exploratory search can be considered as a vector of *m* dimension. Denote the training set of *n* users with each user belonging to one class of people with certain information need:

$$X_{n\times m}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}=\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}^{\mathbf{T}}$$

where x_{ij} is the contribution of the i^{th} user to the j^{th} query or URL, which is a normalized linear combination of the number of times the user has generated queries I_{ij} and the number of times the user has clicked URLs C_{ij}:

$$x_{ij}=\frac{\rho\cdot I_{ij}+(1-\rho)\cdot C_{ij}}{\sum_j I_{ij}+C_{ij}}$$

Defining $\boldsymbol{\mu}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\Phi(\mathbf{x}_i)=\mathbf{x}_i-\boldsymbol{\mu}$ is the mapping function from high-dimensional input data space to a certain feature space, the covariance matrix for this dataset can be calculated as:

$$\begin{aligned} C&=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^{\mathbf{T}} \\ &=\frac{1}{n}\sum_{i=1}^{n}\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^{\mathbf{T}}=\frac{1}{n}\boldsymbol{\Phi}^{\mathbf{T}}\boldsymbol{\Phi} \end{aligned} \tag{4}$$

where $\mathbf{\Phi} = [\Phi(\mathbf{x}_1)^{\mathbf{T}}, \Phi(\mathbf{x}_2)^{\mathbf{T}}, \cdots \Phi(\mathbf{x}_n)^{\mathbf{T}}]^{\mathbf{T}}$.

Then, we have to solve the eigenvalue equation: $\lambda \mathbf{v} = C\mathbf{v}$

where λ is the eigenvalue and $\mathbf{v}$ is the corresponding eigenvector. Since all solutions $\mathbf{v}$ with $\lambda \neq 0$ lie within the span of $\{\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \cdots \Phi(\mathbf{x}_n)\}$ [149], we may consider the following equivalent problem:

$$\lambda \mathbf{\Phi v} = C\mathbf{\Phi v}, \qquad (5)$$

The eigenvector, $\mathbf{v}$, can be expressed in terms of an n-dimensional coefficient column vector $\boldsymbol{\alpha}$ as $\mathbf{v} = \mathbf{\Phi}^{\mathbf{T}}\boldsymbol{\alpha}$. Combining this with (4) and (5) and defining a kernel matrix with n rows and n columns $\mathbf{K}$ by $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathbf{T}}$ will lead to $n\lambda\mathbf{K}\boldsymbol{\alpha} = \mathbf{K}^2\boldsymbol{\alpha}$. To obtain the solution we can solve the kernel eigenvalue problem [149]:

$$n\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha}. \qquad (6)$$

Let $U = (\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_r)$ be the r eigenvectors corresponding to the r largest eigenvalues. Each user contributes to each eigenvector so that a sort of ghostly queries and URLs called *social interest* can be formed. Thus, for a set of original queries or URLs X their corresponding social-interest-based feature Y can be obtained by projecting X into the social interest space as:

$$\mathbf{y}_i = U^{\mathbf{T}}(\mathbf{x}_i - \boldsymbol{\mu}) = U^{\mathbf{T}}\Phi(\mathbf{x}_i)$$

5.4 Social-interest-directed Query Suggestion and Epistemology Retrieval

As our objective is to satisfy users with diverse information needs, we now perform the RWR on the query-URL bipartite graph utilizing the discovered social interest to generate the diversified query suggestion and retrieve epistemologies containing diverse knowledge sets.

5.4.1 Ranking with Interest Measurement

Given a query-URL bipartite graph $G = (Q \bigcup D, E)$, let S be a subset of the vertex set, representing s queries and URLs related to an EIS task, and the user interest measurement $M(i, j \mid S)$ between vertex i and vertex j is the expected social interest of all k random walk paths starting from j to i. It can be calculated using the social-interest-directed feature Y as follows:

$$M(i, j \mid S) = \sum_k \mathbf{y}_k = \sum_k U^{\mathrm{T}} \Phi(\mathbf{x}_k) \qquad (7)$$

$\mathbf{x}_k = [x_{k1}, x_{k2}, ..., x_{ks}]$, and $\begin{matrix} \forall\, m, n \in S, & x_{mn} = R_{mn} \\ \exists\, m, n \in S, & m = i,\ n = j \end{matrix}$.

where R_{mn} is the relevance score or transition probability of vertex m in relation to the starting vertex n.

We can analogize each random walk path on the bipartite graph towards a suggested query or generated result to a user's contribution to the EIS task. The rationale behind Eq.

(7) is: if a user submits queries and clicks URLs iteratively starting from a query vertex *i* to a query or URL vertex *j*, we can deduce the topic the user might be interested in and decide the information need the suggestion of query or result URL *j* is suitable to meet. The user interest measurement from vertex *i* to vertex *j* is determined by the social interest discovered from enormous users, which keeps the subspace that has the largest "variance" so that users with different information needs can be distinguished as much as possible. In the case of the query-URL bipartite graph, given a starting query vertex *i*, if the difference between the interest measurements from *i* to *j* and *i* to *k* is significant, queries or URLs *j* and k must be uncorrelated in the social interest space, which implies that these two vertices probably belong to different topics.

For example, in Figure 5.1, we want to generate the results for query q_1. If we calculate the user interest measurements of all results as: $M(q_1, u_2)$=[0.85, 0.11], $M(q_1, u_3)$=[0.15, 0.69] and $M(q_1, u_4)$= [0.77, 0.02], which indicates that URLs u_2 and u_4 have similar values in all features of the user interest measurement in the social interest space, while URL u_3 has a quite different value. This is because users who submit q_1 and click u_2 have a similar interest with those who submit q_1 and click u_4, while users who submit q_1 and click u_3 have a different interest.

5.4.2 Diversified Query Suggestion and Epistemology Retrieval Algorithm

By unifying the RWR and KPCA in previous sections, Figure 5.2 shows the diversified query suggestion and epistemology retrieval algorithm.

INPUT: Given a query vertex $q \in Q$, the query-URL bipartite graph G, and the transition probability matrix P

PERFORM:

1. Initialize a $1 \times (M+N)$ vector v ($M+N$ is the total number of vertices) for q, with the entry of query vertex q equal to 1 while other entries equal to 0.

2. Run random walk n-steps to get a new vector $v' = vP^n$. Pick the top s vertices with the highest value in vector v' to form a subset S (notice that these queries and URLs are related ones in an exploratory search).

3. For all queries and URLs in the subset S, calculate the covariance matrix as Eq. (4).

4. Solve Eq.(6) and get the r eigenvectors corresponding to the r largest eigenvalues.

5. Calculate the user interest measurement as Eq.(7) for all queries and URLs in the subset S.

6. Move each query with a highest value in one of the r components of the user knowledge measurement from the subset S to the suggested query list.

7. Search the social epistemology repository to get an epistemology that contains URLs with the highest total value in one of the r components of the user interest measurement, and add it to the generated epistemology list.

8. Repeat step 6 and 7 until Z queries are selected and E epistemologies are retrieved.

OUTPUT: Return a ranked list of Z suggested queries and E retrieved epistemologies for query q

Figure 5.2 Diversified Query Suggestion and Epistemology Retrieval Algorithm

5.5 Experimental Validation

In order to evaluate the effectiveness and performance of the proposed social-interest-directed technique for query suggestion and epistemology retrieval, we will compare it with several other methods for EIS tasks through an experimental validation. We define the following task: Given a query and a query-URL bipartite graph, the system has to identify a list of queries and results that are not only most relevant to the given query, but

also diverse to satisfy users with different information needs. We will first introduce the experimental setup, including the dataset, baselines and evaluation metrics, and then present the experimental results.

5.5.1 Dataset and Baselines

We selected the AOL query logs [123] as the base of our experiments. In Chapter 3, we have built up the initial epistemology repository for our system by importing all exploratory searches derived from the AOL query logs. The data collection consists of 796, 735 unique queries and 1, 325, 341 unique URLs after cleaning the raw data. We then get a total of 2, 122, 076 vertices and 5, 642, 820 edges in the query-URL bipartite graph. Moreover, taken as a whole, there are 262, 353 unique terms that appear in all the queries in this data collection.

In order to compare our method with other methods, we randomly select 100 distinct queries from the data collection, and generate a list of suggested queries as well as epistemologies retrieved correspondingly using the proposed and baseline methods.

The epistemology retrieval results of our approach are compared with the following methods, while the query suggestion evaluation is provided in Appendix E:

- Max-sum Diversification (MSD) method [53], which combines the sums of the relevance and diversity measure for the result set as a weighted sum;
- Max-min Diversification (MMD) method [53], which targets at maximizing the sum of the minimum relevance and minimum dissimilarity within the result set;

- Categorical Diversification (CD) method [1], which classifies the queries and documents to categories based on the Open Directory Project (ODP) taxonomy and diversifies the result set to cover all categories.

5.5.2 Epistemology Retrieval Evaluation

We conduct both evaluations to assess the relevance and diversity of retrieved epistemologies respectively, and a comprehensive evaluation to assess relevance and diversity tradeoff in the result sets.

5.5.2.1 Relevance Assessment

First, we evaluate the relevance of retrieved epistemologies. As the objective of diversification is to satisfy users with different information needs, we assess the results' relevance to the user intent, rather than to the query literally.

There are several classical IR metrics widely used for measuring the results relevance, such as MAP and NDCG [102]. However, since the diversification is based on the original results from SEs, we are interested in whether the top documents of the new results are relevant to certain user information needs in general (the diversification has not reduced the relevance by promoting documents irrelevant to any user intent). Therefore, we adopt the Mean Reciprocal Rank (MRR) [102] as the metric for the relevance assessment. MRR is based on the inverse position of the topmost relevant document in the result list, which is calculated as:

$$MRR = \frac{1}{|Q|}\sum_{q \in Q}\frac{1}{r_q}$$

where Q is a set of queries, r_q is the rank of the first relevant document for query q.

For an ambiguous query in the test set, we expect that no matter what the real user information need is, the first epistemology relevant to the user intent could be ranked high. Hence a higher MRR score means the better performance in the relevance assessment.

5.5.2.2 Diversity Assessment

Second, the diversity of retrieved epistemologies is evaluated. To make the diversity assessment, we compute the subtopic recall [196] with the help of *Google Directory*.

Subtopic recall can be used to measure how many concepts are covered by the top results. Consider a query q with m subtopics (concepts) ST_1, ST_2,..., ST_m, and a ranking d_1, d_2,..., d_n of n documents. Let $subtopic(d_i)$ be the set of subtopics to which d_i is relevant. S-Recall at rank k is defined as the percentage of subtopics covered by the first k documents,

i.e., $S-rec@k = \frac{1}{m}\left|\bigcup_{i=1}^{k} subtopic(d_i)\right|$

To find out which subtopic an epistemology is related to, we can use *Google Directory* as a notion of the corresponding category for each search result. While processing an ambiguous query in the test set, we looked up the returned epistemologies' URLs in *Google Directory* and tagged them with the category names under which each URL was listed. Intuitively, if the results covered all possible concepts associated to the query, the value of S-recall equals to 1. Therefore, larger S-recall value means covering more query concepts, or the better performance in the relevance assessment. We reported the S-recall

from S-rec@1 to S-rec@10, and take the average over all the 150 distinct queries in our experiments.

5.5.2.3 Balance of Relevance and Diversity

Finally, we evaluate the effectiveness of results diversification by taking both the relevance and diversity into account. To measure the performance in achieving the balance of these two aspects, the official metric α-NDCG [37], which combines relevance and diversity, is adopted.

The α-normalized discounted cumulative gain (α-NDCG) is an adaptation of the well-known NDCG [102] metric, which is calculated as:

$$NDCG(k) = \frac{1}{Z_k} \sum_{p=1}^{k} \frac{G[p]}{\log(1+p)}$$

where k is a particular top rank position, and Z_k is a normalization factor derived from a perfect ranking algorithm.

For traditional NDCG, the gain vector $G[p]$ is the graded relevance of the result at position p. Let $J(d) = 1$ if the assessor has judged that document d is relevant, and $J(d) = 0$ if not, $G[p] = J(d_p)$.

A parameter α between 0 and 1 balances relevance and diversity in α-NDCG. Consider a query associated with m concepts, let $J(d, i) = 1$ if the assessor has judged that d contains concept Ci, and $J(d, i) = 0$ if not, $G[p]$ is calculated as:

$$G[p] = \sum_{i=1}^{m} J(d_p, i)(1-\alpha)^{r_{i,p-1}}$$, where: $$r_{i,p-1} = \sum_{j=1}^{p-1} J(d_j, i)$$

When α = 0, α-NDCG is equivalent to traditional NDCG. The larger the value of α, diversity is rewarded more over relevance, and vice versa. We fixed α as 0.5 for a balance between relevance and diversity. We reported the α-NDCG values from α-NDCG @5 to α-NDCG @30 of the result sets for all of the 150 distinct queries in the testing dataset.

5.5.2.4 Results

We want to validate whether the social-interest-directed diversification (*SID*) method could retrieve epistemologies that are more diverse but still relevant, as compared to other methods.

Table 5.1 MRR of the four methods

Methods	MRR(Overall)	MRR(Excluding popular user intents)
MSD	0.667	0.141
MMD	0.679	0.159
CD	0.703	0.194
SID	0.711	0.208

The results of the relevance assessment are presented in Table 5.1. It can be seen that there is no significant difference in the overall MRR of the four methods. This is mainly due to the distribution of user intents, since most users are looking for popular information (e.g., Amazon.com), which has already been on the top of the results. However, as the purpose of results diversification is to satisfy the long tail user intents,

we compare the performance of four methods by excluding the popular user intents. The third column in Table 5.2 is the MRR value after removing the popular searches (86 of the 150 quires in the test set) where the first result returned by the SE is relevant to the user intent. The results demonstrate that our method performs best in returning the first relevant documents as early as possible, as it can better capture the conceptual relevance between the query and results by incorporating the social interest.

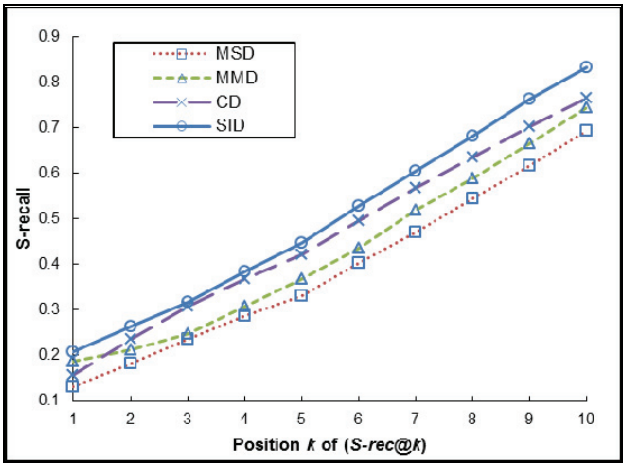

Figure 5.3 S-rec@*k* of the four methods for diversity assessment

Figure 5.3 illustrates the S-recall values of the diversity evaluation from S-rec@1 to S-rec@10. Similarly, *SID* provides better results as compared to content-level approaches *MSD* and *MMD* when measuring the subtopic coverage of the results. Moreover, *SID* also achieves significant improvements over the other concept-level approach *CD* based on *ODP*, since *ODP* is not complete enough to categorize the semantically related queries and documents, which can be better classified through mining the social interest.

Figure 5.4 shows the experimental results for the balance of relevance and diversity, and the evaluation is performed on our method and the three baselines. It plots the results of this evaluation from α-NDCG @5 to α-NDCG @30 for these methods. The performance

of *MSD* is the worst one in the four methods, as *MSD* simply provides more diverse results based on the retrieved documents without considering the relevance. *MMD* gets better performance than *MSD*, but its aim is to maximize the dissimilarity favors rather than satisfying different user information needs, hence the results could not maximize the coverage of the concepts associated with an ambiguous query, while still being as relevant as possible to the user intent (See Table 5.1 and Figure 5.3). *CD* leverages the categorical information from the extra *ODP* taxonomy to cover a wider range of concepts of the query, and thus boosts the long tail user intents for results generation and balances the relevance and diversity. Therefore, it obtains a better performance than both *MSD* and *MMD*. However, among all these four methods, *SID* obtains the best performance as it explicitly addresses the problem for the balance of relevance and diversity by introducing the social interest space to replace the query and result space, hence the results can be organized based on the user intents embedded into different concepts of the query. Further, the experimental results of this evaluation also demonstrate the effectiveness of our proposed diversified results generation algorithm.

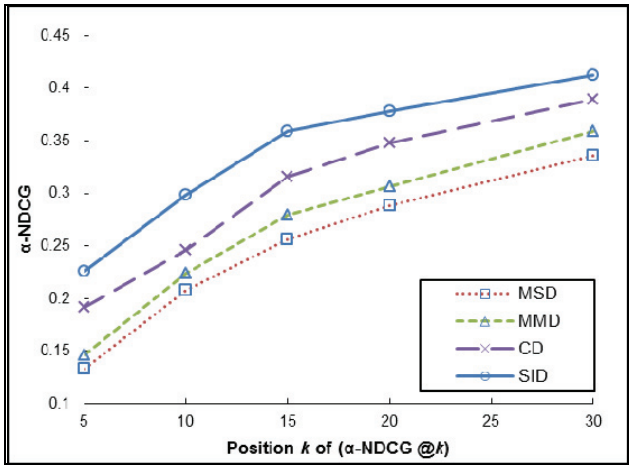

Figure 5.4 α-NDCG *@k* of the four methods for the balance of relevance and diversity

5.6 Summary

In this chapter, we have presented a novel social-interest-directed technique to addressing the critical epistemology retrieval issue in the proposed social search solution. Since users have difficulties in formulating proper keywords in the EIS process, query suggestion and result diversification techniques are used to tackle the epistemology retrieval issue. Unlike previous techniques revolving around mining existing queries that are most similar to a given query or providing most dissimilar results for a given query, the proposed technique can suggest highly diverse queries that are yet closely related to a given query, and provide diverse epistemologies for users with vague information needs.

Social interest is discovered by employing KPCA on the related queries and epistemologies contributed by enormous users. To suggest queries and retrieve epistemologies that satisfy both the relevance and diversity, an algorithm is developed to incorporate the social interest with a random walk on the query-URL bipartite graph. We also demonstrate that our solution outperforms other methods in terms of suggesting queries and providing results for users with vague information needs through experimental evaluations. The suggested queries and retrieved results are diversified to satisfy users with different information needs, while they are still relevant to the given query.

This work could be further extended by exploring other statistical data analysis methods, such as factor analysis, to improve the performance of mining social interest in the future.

Chapter 6

Epistemology Editing

Existing search services only acquire information that is related to users' needs rather than the information that they exactly want. One of the advantages of our epistemology-based social search solution over conventional SEs is that the epistemology can be further edited and refined to help prosumers achieve their final search goal. The exponential growth of UGC on the Web also convinces us that it is possible and necessary for prosumers to work on the epistemologies collaboratively. In this chapter, we address the epistemology editing and refining issue through an approach that can help users acquire the information provided on demand. This approach integrates social networking into the EPISOSE framework, where users in a social search community contribute epistemologies and form social networks from their search activities. Information is provided on demand through a consumer-led interactive search process, where invited information providers from relevant social networks jointly edit and refine the epistemology to meet the consumer's needs. We applied this approach in the prototype social search system that uses an epistemology structure to represent a consumer's information needs and lead an interactive search process. Preliminary user feedback has shown that this approach can improve search efficiency and quality in various social search situations, and is particularly effective for a consumer to acquire diverse but coherent information through the epistemology editing and refining.

6.1 Introduction

The technological progress since Web 2.0 has made it free and easy for anyone to publish any content at any time from any place, which has resulted in an explosion of UGC. UGC refers to different kinds of media content on the Web created and published by the Internet users voluntarily, such as wikis, videos, or restaurant ratings. In recent years, with the popularity of social networking and micro-blogging sites, UGC has experienced an even rapid growth on Internet, partially because it can be obtained free of charge. For instance, users can search for latest tweets posted on *Twitter*, or friends' newsfeed on *Facebook*, or existing epistemologies on a topic of interest contributed by prosumers in *Baijia*, with their built-in search services.

However, while the great mass of UGC does not contain useful information for most people on the Web, it is getting more difficult to find relevant information from the increasing volume of UGC even with the help of SEs. Especially, general-purpose SEs (e.g. *Google* or *Bing*) are unable to follow after UGC well because of the crawl-delay, i.e., the crawlers need to take significantly long time to discover and index each new page. Moreover, they are unable to properly display UGC to the users because the content would not be displayed on the top of the search results due to low PageRank [121].

Specialized UGC search services, such as *Twitter Search* or *Google Real-time Search* [151], partially solve general-purpose SEs' problems with UGC by using user-provided keywords to constantly keep up with UGC updates on a certain topic/person, e.g. activities, blog posts, newsfeeds, and tweets, in social media.

Nevertheless, these search services are still similar to general-purpose SEs that locate resources on the Web according to users' queries. Due to the vocabulary problem [47], i.e., one topic can be described by different people using different vocabulary, a vague search goal is hard to formulate with proper keywords. If a query does not retrieve the information that a consumer (i.e. an information seeker) exactly wants, she/he has to amend the query using different keywords. If such an iterative trial-and-error process does not end up with what she/he wants, the information is deemed non-existent and she/he would give up and try again some time later.

Recently, social search has been adopted to address these difficulties by utilizing the wisdom of crowds [162]: as many people search for the same or similar information, reusing and refining others' successful searches are pragmatic solutions. It has changed a search process from an individual activity to a social one. Social search has been proved effective in information filtering (e.g., understanding users' search goals) and helpful in information post-processing (e.g., annotating search results).

However, current social search systems are still inadequate for supporting effective UGC search. This is mainly due to the fact that information providers are not actively involved in search processes to help information consumers easily access the information they have provided. According to Heymann's work [66], although some information providers publish information (e.g., answers to a question) not currently available at other sources, they need to make extra efforts to make the information reach as many information consumers as possible (e.g., make it searchable by common SEs).

Specifically, previous work has generally omitted the characteristic of UGC. The advent of UGC has transformed the Web from monologues to dialogues, since it encourages one to publish their own content as well as to comment on others'. Such transformation is similar in significance to the evolution of conventional media, e.g., from broadcast TV to Video on Demand (VOD). Furthermore, the creation of UGC outside the professional routines and practices has transformed the Internet users from primarily content consumers to providers, because facilities have been provided for amateurs to publish their own content.

In general, the two-way content delivery between ordinary Internet users (in contrast to the one-way delivery from professional Web designers to a passive unseen audience) implies that the information providers could be able to prepare or update their content to meet the needs of the information consumers. Especially for the content provided by numerous users in the popular social media platforms such as *Twitter* and *Facebook*, or a social search community, if the providers who have published some information are informed of what else consumers are looking for, they would be able to generate such content. For example, in the study of Teevan et al. [166], a lot of people published tweets about the movie *New Moon*. However, there are many questions about whether the movie was worth seeing. If the provider who published content related to "new moon" were informed of the queries such as "recommendation", she/he would publish a new tweet, e.g., "I recommend the movie new moon that is as good as the movie twilight".

Therefore, based on the social epistemology concept, we propose the Information Provision on Demand (IPOD) approach for users to acquire non-existent information

though epistemology editing and refining. Social networks are established for prosumers with same or similar interest, and social network analysis is performed to locate appropriate prosumers for providing information on demand. Consequently the consumers can acquire the non-existent information they exactly want in a consumer-led interactive search process. More specifically, the IPOD approach can be summarized as follows:

First, social networks are formed in the social search community for prosumers to effectively share, reuse and refine the epistemologies. Second, an information consumer can lead an interactive search process involving prosumers in the relevant social networks in the community through a pre-structured epistemology representing her/his personalized information needs. Third, the consumer can interact with the providers to clarify, comment on, refine, or request more information on the infused epistemology.

The IPOD approach has been applied to the implementation of a UGC search service based on the EPISOSE framework. The framework builds social networks based on the search epistemologies in the community, and further exploits these social networks to identify proper information providers for conducting effective consumer-led interactive search tasks.

We have constructed the epistemology repository and social networks of the community based on automatically generated epistemologies in the prototype system *Baijia*. For the purpose of conducting internal usability testing, we have also implemented a micro-blogging system for information providers to publish UGC. Preliminary user feedback

has shown that this approach is particularly effective for a consumer to acquire a structured knowledge unit consisting of diverse but coherent information.

6.2 Related Work

Recent years have witnessed a phenomenal growth of UGC on the Web, mainly owing to the increasing popularity of various micro-blogging systems and social networking sites. For example, *Twitter* users can publish real-time topical news [125]. However, without a proper SE, the vast majority of such content is only visible to certain social networking contacts rather than reaching general public.

While the original search scheme generally does not accommodate UGC search, studies on social activities and human factors on Web search are now in the ascendant. Collaborative tagging systems exemplified by bookmark sharing (Delicious[13]), photo sharing (Flickr[14]), and video sharing (YouTube), allow users create various tags to make "shared" easy for retrieval. The community search assistant [50] enables a community of searchers to search in a collaborative fashion by using query recommendation based on a graph where related queries are connected.

There are some approaches that also exhibit some relevance to the consumer-led or interactive process, although they are not specially catered for UGC search. For example, QA systems such as *Yahoo! Answers* can be regarded as a consumer-led process as a user

[13] http://www.delicious.com
[14] http://www.flickr.com

directly posts her/his question on the system in order to lead provider(s) to prepare their answers accordingly. Some personalized or vertical search systems based on interest-based ranking [188] or topic-specified crawling [109] can also be regarded as a consumer-led process as a user's profile (instead of a topic query) is used to lead the search. Some social search systems can also acquire information from the providers based on the social interaction in search processes [46][80]. For example, *Google Aardvark*[15] provides various communication tools (e.g. instant message, email, etc.) for a user to interact with her/his friends (information providers) during a search process. Consumers may also use a general-purpose SE to find specific websites, where they can directly interact with providers through online chat.

The nature of epistemology editing and refining depends on the specific problem context and our unique problem context is to meet an individual consumer's personalized and diverse information needs through the UGC. That is, a consumer could guide the search process to meet her/his personalized information needs, interact with information provider(s) to polish acquired information towards her/his needs, and coordinate multiple providers who cooperatively contribute to the consumer-defined epistemology regarding a specific topic. Therefore, compared to the related work, our approach contains a nonspecific process: for providers, the content is not just published for a specific question or consumer and other consumers are difficult to acquire the content; for consumers, they do not need to find out specific information source first and only acquire information from that source.

---

[15] http://www.vark.com

Most existing research works on social networks are focused on looking for people with specified names [152][112]. Although users can benefit from the social networks since they can connect with experts in a given subject area when they need advice or help, current social search systems [69] try to help users with their problems by discovering and enquiring experts only in their existing explicit social networks (e.g. social networks constructed in *Facebook*). However, new social networks can be formed during social search processes in our approach, as it is possible that people with same interest or similar information needs might network with each other. For example, some previous system, such as Maze [194], allows users to create friends in the file sharing network. Moreover, based on the users' activities and epistemologies contributed, we can not only discover their interests and information needs but also identify their expertise.

Our approach emphasizes the role of the social network of prosumers and their collaboration in epistemology editing and refining. Moreover, a prosumer's social network is established based on the epistemologies in the search community by clustering prosumers with same or similar interest, and the further analysis of the social network structure with artificial intelligence techniques can help discover providers who will or may generate information related to a consumer's information needs.

6.3 Information Provision on Demand

6.3.1 Leveraging Social Networks for Epistemology-based Social Search

In social search systems designed and implemented with the guidance of the EPISOSE framework, prosumers can be engaged in reaching out to old friends and meeting new

ones who share their interests. They may also edit, or refine epistemologies, and communicate with others directly. However, prosumers' searches can benefit from the building of social networks, and proper strategies can be adopted in this framework for effectively locating expertise in social networks.

It is a common phenomenon that users would be likely to look for help from others while they are conducting search tasks, if they are unfamiliar with the subject area of that task. If users are not sure about what they are looking for, seeking advices from experts in the right areas is always a good option. In our approach, prosumers can find people with same interests or similar information needs from social epistemologies and thence build a social network with them. In other words, the social network is constructed from social search activities, and the constructed social network will improve social search activities in turn.

The technical details of the interaction of social network and social search in our approach are given in Appendix F. In our approach, prosumers are categorized to different clusters by the traditional *k*-means clustering algorithm, which partitions the prosumers into *k* sets in a way that minimizes the variance within each group. It should be noted that users usually have multiple information needs and interests, which will result in grouping a user into multiple clusters. In such case, algorithms such as fuzzy *C*-means can be applied. The representation of clusters in fuzzy set makes it possible for a prosumer to belong to multiple groups with a degree of membership between 0 and 1.

Reliable prosumers can be located to improve the search experience by incorporating social networks into the IPOD approach. Moreover, the fact considering prosumer

clustering in social networks also suggests that it has the potential to defend against several attacks or gaming in rating systems. By relying on connections in the social network, it is possible to eliminate many types of attacks that rely on creating multiple accounts, such as Sybil attack [43]. While these accounts could all connect to one another with high trust, they would only be clustered with "good" prosumers if some of these good prosumers assigned them high ratings as well.

6.3.2 Consumer-Led Epistemology-Mediated Interactive Search

We have provided a generic social search framework based on epistemologies contributed by numerous prosumers in a social search community, and built the social networks in the community based on the epistemology-mediated user correlation. As we have located the prosumers with highest reputation and expertise in a consumer's social network that probably can provide satisfactory information for the consumer, we can now support the consumer-led interactive search in the IPOD approach. It should be clarified that our approach is independent of existing social networks, e.g., *Facebook*. Real-world social networks can be adopted to boost or augment the consumer-led interactive search process, e.g., only interacting with familiar friends.

6.3.2.1 Initialization and Test

We first constructed the initial epistemology repository and social networks of the community based on automatically generated epistemologies. We also used the initial repository to test how much the IPOD approach can outperform a conventional SE while integrating social networking into the epistemology-based social search.

We built up the initial epistemology repository by importing all search processes derived from the AOL query logs as described in Chapter 3. We also adopt the same metrics used in Chapter 3, including MAP, and NDCG, to compare the performance of the following approaches:

The first approach is the AOL SE, where the results are derived from the original data;

The second approach is IPOD without social networking, where the search results are derived from all relevant epistemologies;

The third approach is IPOD with social networking, where the search results are derived from epistemologies contributed by similar users clustered in the social network.

Figure 6.1 shows the MAP scores of the IPOD approach as compared to those of the AOL SE. As expected, the MAP score has been improved significantly with IPOD approach and a growth in the search precision is shown when searching with IPOD while social networking is integrated. The result looks quite reasonable from the way our test was conducted, as while more searches are imported and more users are included in the social network, a user who involves in a search process will have a higher probability to get relevant search epistemologies and more users with similar interest whose epistemologies can be utilized to satisfy the user's information needs.

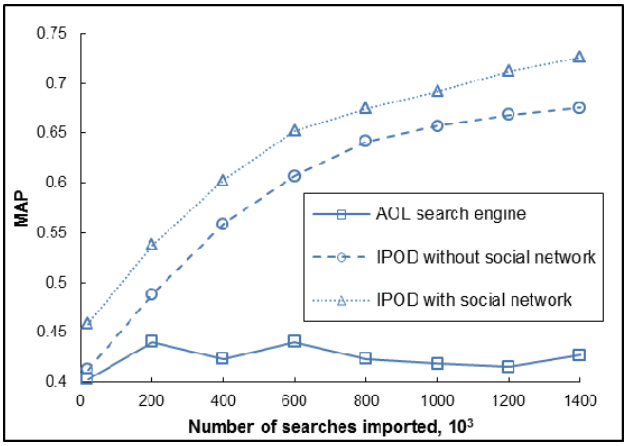

Figure 6.1 MAP scores of the IPOD approach and the AOL SE

Figure 6.2 shows the NDCG@10 of the IPOD approach and the AOL SE. We can observe that the ranking algorithm adopted in IPOD with social networking outperforms the algorithm without social networking and the algorithm adopted in AOL SE.

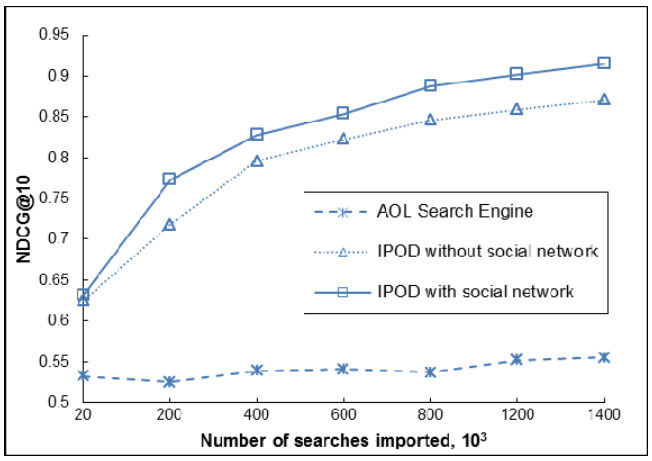

Figure 6.2 NDCG@10 of the IPOD approach and the AOL SE

Based on the testing results, we confirmed that integrating social networking into the epistemology-based social search can improve the search quality and the algorithm adopted in IPOD is effective to identify relevant prosumers for the consumer-led interactive search.

6.3.2.2 Consumer-led Interactive Search Overview

A schematic architecture of the consumer-led epistemology-mediated interactive search approach is shown in Figure 6.3, which sketches key sub-components of the *Epistemology Editing & Refining* component in the EPISOSE framework.

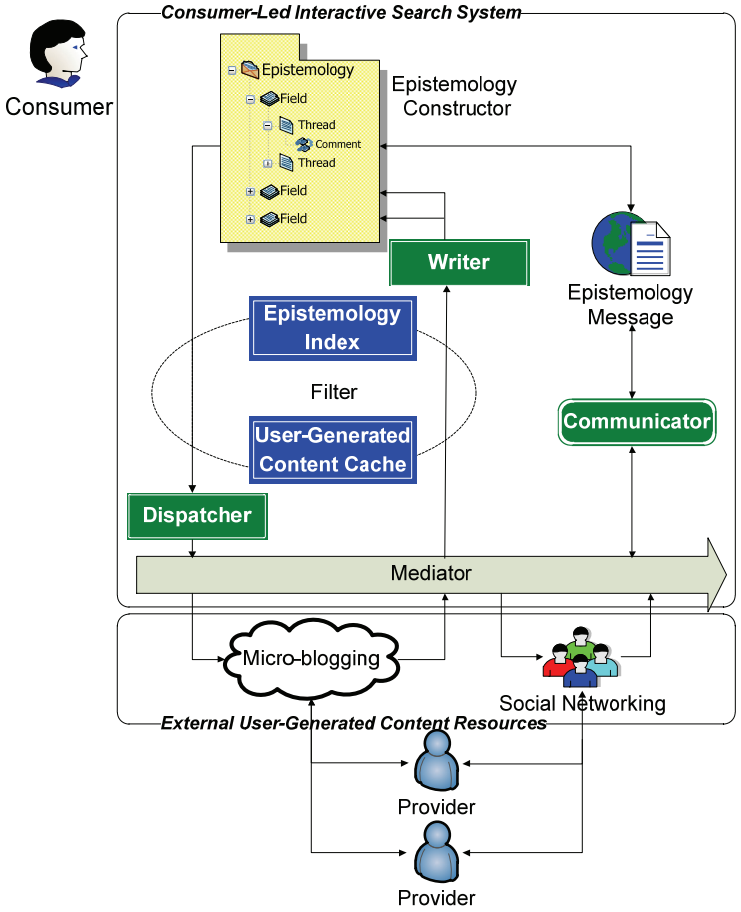

Figure 6.3 The schematic architecture of consumer-led epistemology-mediated interactive search

Epistemology Constructor – for a consumer to create a pre-structured epistemology depicting a blueprint to lead an interactive search process. The structure of epistemology will be discussed shortly.

Epistemology Index - for the *Filter* component to quickly retrieve the epistemology that can be filled out with relevant UGC.

User-Generated Content Cache – for the *Filter* component to discover cached UGC from external social media systems that is related to an epistemology created by the consumer.

Filter – a core component that retrieves *User-Generated Content Cache* and *Epistemology Index* in order to find matches between relevant fields of an epistemology and cached UGC. In addition, it explores the social networks formed in the system to locate the providers whose content matches the epistemology, and collects the contact information of the providers so that the *Dispatcher* component can dispatch the epistemology to them in order to lead the search process and facilitate interaction between the consumer and these providers via the *Communicator* component. It also passes on the matched UGC from the cache to the *Writer* component, which will then fill them into the relevant fields of the epistemology.

Mediator – the bridge between the consumer-led interactive search system and external user-generated resources provided by micro-blogging users or social networking friends.

6.3.2.3 The Pre-structured Epistemology

Epistemology is structured hierarchically, as shown in Figure 6.4. An epistemology, which describes a consumer's information needs for a specific topic, consists of a list of separate but inter-related *fields* (a.k.a. sub-topics) and each field is composed of a set of independent or inter-related *threads* (for interaction between consumer and providers).

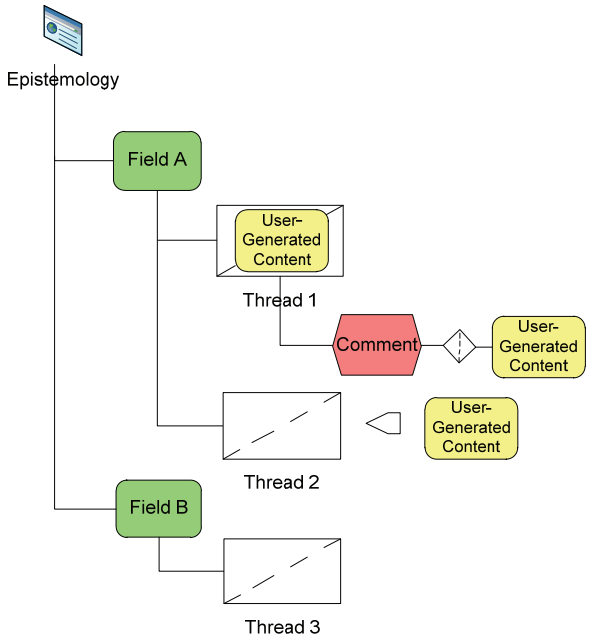

Figure 6.4 The structure of the epistemology

Field is the working unit in an epistemology. Each field is tagged with a set of sensible keywords that will be used by the *Filter* component to precisely match with cached UGC. Matched content will be filled into the field by the *Writer* component. *Thread* is the interaction and cooperation unit. The consumer can actively interact with information provider(s) that have contributed to a specific field by commenting on their input in order to clarify doubts, correct errors, or polish the results. Multiple information providers may jointly input their content to the different threads of the same field, or even different threads of different fields, no matter whether they are aware of their cooperation.

The epistemology structure is obviously advantageous in the case that a consumer wants to search for updates on a topic that cannot be simply described by a few keywords. With an ordinary UGC search services, the consumer has to generate multiple queries using different keywords on multiple instances of a Web browser because using all keywords in a single query is likely to get no matched result at all.

In contrast, with the consumer-led interactive search system using structured epistemology, multiple queries using different keywords can be generated simultaneously and matched results will be filled into the corresponding fields of the epistemology automatically and simultaneously. More importantly, the generation of multiple queries is completely transparent to the consumer, who only sees what she/he wants to know have been filled with results and who may wish to take the chance to interact with the information providers just for the sake of polishing the results. Furthermore, the consumer can always interact with providers to clarify doubts, correct errors, or polish the results.

6.3.3 The Consumer-Led Interactive Search System

We have applied the IPOD approach to the *Baijia* prototype system. We have also implemented a micro-blogging system for information providers to publish content for the purpose of conducting internal usability testing. We will discuss some user interface features of the systems in this section. Note that the external micro-blogging system is connected to the search system via the *Mediator* component within the system by using the micro-blogging system's API. Therefore it is also possible to connect the search system with various external social media system via their APIs, e.g. *Twitter*, *Facebook*, and so on.

6.3.3.1 Epistemology Constructor and Filter Interface

In our system, a consumer can create a pre-structured epistemology and use queries and related phases to present the information needs as far as possible. As shown in Figure 6.5,

the consumer is seeking for information about the "World Cup Final". She/he would like to acquire information about both sides of the match, therefore a pre-structured epistemology is created with two fields: "Spain in World Cup", and "Netherlands in World Cup". Each field will be a container for related UGC.

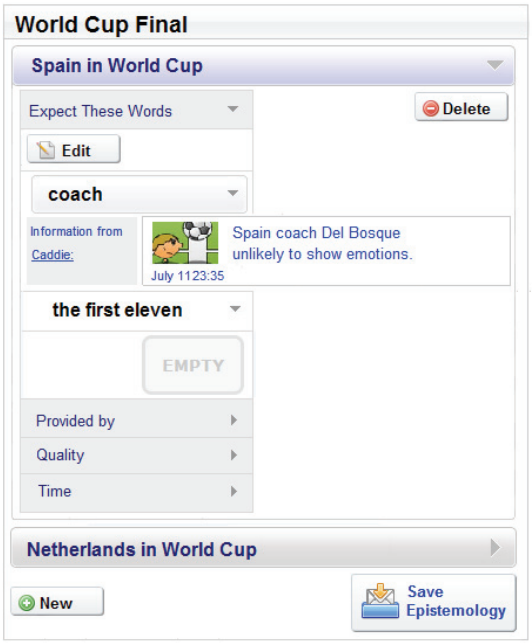

Figure 6.5 The epistemology constructor and filter interface

In the first field, the consumer is interested in the activities of Spain in the World Cup. Therefore she/he can define some rules to filter the UGC that meets the information needs. For example: "Expect these words" such as "coach", "the first eleven", and so on, in the UGC. The search service will generate the conditional expression "include (Spain in World Cup) AND (coach OR (the first eleven))" in the epistemology index. Once there is information in the *User-Generated Content Cache* that satisfies the condition, e.g. "Spain coach Del Bosque …", it will be inserted into the field immediately. At the same time, the inter-related phase "the first eleven" will be dispatched to suggest the providers to publish relevant content.

The consumer can also filter the information with specified providers, e.g. the content from "Messi"; or filter according to the expertise levels of providers, e.g. grade, ranking by viewers, the number of the followers; or filter according to time, e.g. "first publishing" or "last updating".

6.3.3.2 Micro-blogging Interface

Figure 6.6 shows the interface of the micro-blogging system connected with the consumer-led interactive search system. Providers can easily publish content through such a *Twitter*-like interface.

Figure 6.6 The micro-blogging interface

While a provider types a piece of content, relevant epistemologies dispatched from the consumer-led interactive search system are being filtered in and displayed so as to suggest her/him to give as much relevant information as needed by consumers.

For example, while a provider types *"Spain sweat over Villa injury"*, the micro-blogging system immediately sends it to the consumer-led interactive search system, which in turn filters all pre-structured epistemologies to find those asking for such information, e.g. the epistemologies titled "*Spain World Cup"* and *"Expect these words: injury, influence"*.

The consumer-led interactive search system then dispatches these matched epistemologies to the micro-blogging system and posts a micro-blog there, e.g. "Currently interested information about #Spain #injury: #substitution(12) #squad(5) #influence(1)". Such information will suggest the provider to publish the content that is most interested (i.e. many consumers are interested in the substitution for Villa) based on the number following each tag indicates, which is calculated as:

$$N = N_{epi} - N_{rec}$$

where N_{epi} is the number of epistemologies are expecting content tagged by that keyword, and N_{rec} is the number of providers who has published any content tagged by that keyword.

6.3.3.3 Writer and Communicator Interface

At the same time, the writer component will update the pre-structured epistemologies that are requiring such information by filling a user-defined field with the relevant content identified by the *Filter* component, as shown in Figure 6.7. Relevant content is identified using the traditional approach in IR research: term frequency – inverse document frequency (TF-IDF) [143].

For each term t_i of the query in the user-defined field, $tf(t_i)$ is the frequency of t_i in UGC, and we calculate the IDF in the form:

$$idf(t_i) = \log \frac{N}{df(t_i)}$$

where N is the number of all UGC and $df(t_i)$ is number of UGC where the term t_i appears.

Then: $tf\text{-}idf(t_i) = tf(t_i) \times idf(t_i)$

Figure 6.7 The Writer and Communicator interface with the epistemology

However, for some UGC such as tweet, since users tend to remove word redundancy from a tweet to save space, seldom terms are repeated in a tweet. In that case, TF-IDF is essentially just the IDF term.

The content and comments are organized based on different providers. Therefore those providers can work jointly and concurrently while the structure of the epistemology remains unchanged. That could make the epistemology more readable than search results ordered by time.

Thereafter, the consumer is able to interact with the provider(s). For example, the consumer can comment on the thread *"Pele: Spain Are Favorites to Win World Cup"* by *"How about the prediction of Octopus Paul"* and the provider can publish another piece of content to update the epistemology accordingly.

6.4 User Feedback

We conducted a preliminary usability testing of the *Baijia* search system in order to: 1) understand whether users (i.e. information consumers and providers) like this new concept and if yes what features they particularly like, 2) investigate what kind of search tasks where the consumer-led interactive search system does better than existing UGC search services, 3) study whether the system is easy-to-use and what special skills users need to use the system, and 4) get some feedback to improve the system.

6.4.1 Participants and Tasks

Ten users participated in this small-scale user research. The users included both men and women, and ages ranged from around 20 to around 50 years old (median=27). Five of them were undergraduate and postgraduate students, three were staffs, and the rest were IT professionals. As we chose *Twitter* as a representative UGC media, eight of them have the experience of using UGC search services and seven of them are current users of *Twitter* and have the experience of publishing content on *Twitter*.

To investigate what kind of search tasks where the consumer-led interactive search system might do better than existing search systems, we deliberately designed two search tasks.

The first scenario was searching for UGC about the *new generation iPhone*. Since the consumers were eager to know more details about that product from a current owner or an expert, it was very likely for the consumer to invite the providers to publish their diverse views to the consumer's information needs and such personal knowledge could not be retrieved through existing search systems.

The second scenario was searching for UGC about the *World Cup Final*. Since the consumers were pretty clear about what they were looking for, e.g. goals, shoots, etc., they could clearly define an epistemology structure to invite the providers to publish content that meets their needs.

The same set of participants did the two scenarios first with *Twitter Search* and with the UGC published on *Twitter* and then with the *Baijia* system and with the UGC published on our micro-blogging system. They were not allowed to use any other communication channels, e.g. phones or instant messengers, except the systems given to them.

At the end of the testing, we interviewed the participants in order to understand their views on the system, in terms of novelty and search results quality, as well as the advantages and limitations of the system.

6.4.2 Scenario 1

In this scenario, seven consumers were interested in buying the new generation iPhone. However, they have been bored by the perpetual advertisements and stereotyped reviews on the Web. Therefore, they turned to search for UGC about the latest and just critiques on the product. Three providers were either the current owners or technical experts of the product. Since the consumers were eager to know more details, and the providers would also like to share their good or bad experience, it was very likely for the consumer and the providers would interact frequently in the search process.

Participants were impressed by the *Baijia*'s ability of allowing consumers to lead the interactive search process, and invite multiple providers to publish their content incorporated in different fields of the epistemology in the search process. Since the consumers would make the decision for buying the product based on the search results, they had to search for advantages and disadvantages reflected by the owners. Most of such information could not be retrieved by *Twitter Search* since many providers generally would not publish their personal views on specific details. *Baijia* was able to suggest them to publish required information through promoting with epistemologies created by consumers on the same topic as the content already generated by the provider, such as "*my latest iPhone*".

It is because the search process is led by the consumer that makes the epistemologies so valuable to a particular consumer for her/his particular information needs. All participants commented on that feature during the interview.

6.4.3 Scenario 2

In this scenario, the participants were situated in context of the 2010 World Cup final. Six consumers couldn't see the live telecast for some reasons, so that they searched information about the match on the web, while the other four participants were designated as providers to publish content. Since the consumers were pretty clear about what they were looking for, e.g. goals, shoots, etc., they could clearly define an epistemology structure to lead the providers to publish content that meets their needs.

All participants agreed that the consumer-led interactive search system brought a novel user experience, both for consumers and providers. As the search was mediated by structured epistemologies, it was particularly efficient for consumers to complete a search task that generates multiple or complex queries; it was also effective for providers to generate high quality content that could meet the more consumers' needs. Further, the quality of search results of the *Baijia* search system was much more satisfied than *Twitter Search*, not only because they could get more information through interaction with the provider, but also because it was really hard to read on some topic in *Twitter Search* which only had a simple single column filled with feed. In contrast, in the structured epistemology created by the consumer, the UGC was catalogued into fields with multiple queries simultaneously and automatically, which was more readable than a sequence of content that updated frequently.

6.5 Summary

Recent years have witnessed a phenomenal growth of UGC on the Web, the epistemology editing and refining component in EPISOSE framework can effectively help get such content across to general public (instead of within certain social networking contacts). General-purpose SEs and specialized UGC search services could not help consumers acquire the information they exactly want. In this chapter, we propose the IPOD approach to help consumers acquire the non-existent information by addressing the epistemology editing and refining issue in the epistemology-based social search solution.

In this approach, information is provided on demand through a consumer-led interactive search process where invited information providers from relevant social networks jointly edit and refine the epistemology on the fly to meet the consumer's needs. We have implemented the approach in the *Baijia* prototype system, where prosumers with the same or similar search interests are clustered in the social networks built based on the epistemologies. The system thereby can identify potential providers in the social network of a consumer, and use a pre-structured epistemology to represent a consumer's information needs so that the prosumers can edit and refine the epistemology in an interactive search process. Initial usability testing of the system has given positive feedback to the approach, which confirms the improvement of the search efficiency and quality in various social search situations. Our system can be further improved by connecting it with public social media systems, e.g. getting UGC through *"Twitter Firehose"* to refine epistemologies.

Chapter 7

Trustworthy Social Networking

In previous chapters, we have shown that social networks can be utilized to improve the quality of social search experience. Trust management is a paramount issue in *Social Networks Building*, which is one of the important functions of the *Epistemology Services* component in the EPISOSE framework. In this chapter, we contribute a novel trust model that allows personalized measures to be naturally established on the objective ground. The key concept of this model is to infer trust by tracing the credit flow within a social network, where the trust between a pair of nodes can be derived from the credit flowing from one node into the other and the relative risk disparity between them. This model, inspired by the physical and mathematical properties and the power flow study in electrical grids, is based on the hypothesis that the credit flow in a social network is fundamentally similar to the power flow in an electrical grid. Experiments with a real-world dataset have proved the hypothesis and the results have shown that the credit-flow-based trust model can derive personalized and more accurate trustworthiness than existing models do.

7.1 Introduction

Recent years have witnessed not only the widespread of various online social networks but also the exponential growth of virtual relationships. Online social networks take

various forms pertaining to how virtual relationships between parties are established. A relationship could arise from online interaction, for example, one making comments on a video posted to YouTube by another, or from online transactions, for example, one purchasing an item from another through eBay, or from friendships inherited from the real world or encountered in the virtual world, for example, friends on Facebook. In our epistemology-based social search solution, prosumers with the same or similar search goals build up trustable social networks to complete their search tasks together.

Trust management becomes a paramount issue in online social networks as the population of social networking users receiving critical information from or making financial transactions with their virtual relations has skyrocketed. In recent years, a number of trust models have been proposed and used for trust management in a variety of social networks. Depending on whether global or local metrics are used to derive trust measures, they can basically be divided into global or local models [201].

A global model adopts a centralized strategy to define a unified measure of trust for all nodes in a social network. It assigns a trustworthiness score to each node and every other node in the network trusts it in a unanimous way that is decided by the score. For example, if Alice has a higher trustworthiness score than Bob does, every other user in the network would trust Alice more than Bob. A global model can be regarded as a centralized reputation model, where each node's reputation is calculated by global metrics that are based on the whole network structure or link analysis. For example, with the PageRank [121] algorithm or its variations [144], Alice would have a good reputation too if nodes linking to her were all reputable.

In contrast, a local model adopts a distributed strategy to define a personalized measure of trust for every individual node in a social network. It assigns multiple trust values to each node, one for every other node in the network and every other node trusts it in such a way that is decided by the corresponding trust value. For example, Alice's trust value towards Carol is 0.6, while Bob's trust value towards Carol is 0.1, indicating that Alice trusts Carol more than Bob does. To calculate the trust value for every pair of nodes in a network, direct trust values between neighboring nodes must be collected from users as additional input parameters and local trust metrics need to propagate direct trust values through the network in a peer-to-peer manner to derive indirect ones. For example, assuming distributed trust propagation along a path is multiplicative, if Alice's trust towards Bob is 0.6 and Bob's trust towards Carol is 0.5, Alice's trust towards Carol may be $0.6 \times 0.5 = 0.3$.

Global trust models are suitable for social networks where explicit personal trust values for all nodes are not available, but interactions between nodes can be used to infer the overall reputation of each node by a centralized authority. As bias is a natural property of trust, global models are too objective in that they do not support personalized measures towards the same node. While local models have addressed this issue as the trust propagation is done by the nodes themselves in a distributed manner in order to calculate personalized trust values for each node, they tend to be too subjective in that they do not consider a social network in its entirety. Apart from that, it is non-trivial to collect accurate personal trust values only based on partial network information.

More importantly, neither global nor local models have taken into the consideration the risk factor of trust management, although risk is intuitively related to trust. For example, Alice happens to own high credit in a particular field but people who gave her credit are novices in that field. In contrast, Bob, who is in the same social network, owns relatively lower credit but people who gave him credit are all experts in the field. In this case, Alice may have a higher risk factor than Bob does and therefore should not necessarily be trusted more than Bob.

In this chapter, we contribute a novel distributed trust model CoreTrust (Credit Over Risk Equals Trust), which allows personalized measures derived from distributed trust propagation to be naturally established on the objective ground. The rationale behind this new model has three folds. First, a node's objective reputation inferred from interactions between nodes in an entire social network is more reliable than a subjective trust value explicitly specified only based on partial network information. Second, since bias is a natural property of trust, it is simplistic to assign a universal reputation to each node and force every one else to trust it in a unanimous way. Last, incorporation of subjective risk assessment between nodes with each node's objective reputation is likely to yield personalized and more accurate trust measures.

The key concept of the model is to infer trust by tracing the credit flow within a social network, where the trust between a pair of nodes can be derived from the credit flowing from one node into the other and the relative risk disparity between them. This credit-flow-based trust model, inspired by the physical and mathematical properties and the power flow study in electrical grids, is based on the hypothesis that the credit flow in a

social network is fundamentally similar to the power flow in an electrical grid. Experiments with a real-world dataset have proved the hypothesis and the results have shown that the credit-flow-based trust model can derive personalized and more accurate trustworthiness than existing models do.

7.2 Related Work

Social computing applications and social networking services such as *Facebook* and *Google+* have attracted a good deal of interest over the last few years. In most social applications or services, including recommender systems and our epistemology-based social search solution, trust management is crucial for someone to determine the trustworthiness of anyone else with whom she/he will exchange information. Especially for those who do not personally know each other in the real world or have no prior direct interaction in the virtual world, trust inference is a critical approach to establishing new trust measures in a social network [100].

7.2.1 Global & Local Trust Models

Trust models that centrally calculate a universal measure of trust for all users in a social network are classified as global, as the objective of these models is to rank all nodes with a global reputation. For example, Kamvar et al. [81] proposed the EigenTrust algorithm to calculate trust rating of each node in a network with a variation of the PageRank algorithm. Richardson et al. [133] described an approach that first finds all paths from a node to every other node, each of which represents an opinion of a statement, and then

aggregates trust values along every path to calculate the final trust value. Guha [58] built a generic trust engine using the TrustRank and DistrustRank algorithms to capture the global trust rankings, which allows people to rate the content and the associated ratings from others. Xiong and Liu [186] presented the PeerTrust model, where the trustworthiness of a peer is calculated as the average feedback weighted by the rankings of the feedback contributors. The PowerTrust [203] and GossipTrust [204] models proposed by Zhou and Hwang use the power-law feedback characteristics and the power of gossip respectively to disseminate feedback and reputation data from which a global trustworthiness value for each peer will be derived. Liu et al. [100] described StereoTrust, which attempts to derive the expected trust by aggregating stereotypes - built on the basis of existing agents' observed membership of particular groups – that match an unknown agent's profile.

In contrast, trust models that calculate personalized trust values for each user in a distributed manner are classified as local, as these models take into account individual bias based on the partial network information. Local trust models exploit a user's personal experience and the web of trust to compute the trust value for every other user in the network. For example, Jøsang [79] introduced subjective logics to assess trust values based on the triplet representation of trust. Similarly, based on the Dempster-Shafer theory of evidence and the explicit notion of uncertainty [150], Yu and Singh [195] developed a heuristic discounting approach that combines the local evidence with the testimonies of others evaluating the trustworthiness. Raph Levin's Advogato project [93] applied the network flow in graph theory to a modified graph, which composes certificates between members to determine a member's trust level and their membership

within a group. Ziegler and Lausen [206] presented the Appleseed algorithm for local group trust computation, which was inspired by propagation of activation over a network like neurons in psychology. Golbeck [52] proposed the TidalTrust model to infer trust value in continuous trust networks. When a node wants to infer the trust rating of a sink node, it first asks its trusted neighbors for the rating of that node, and then calculates a weighted average trust rating of its neighbors to the sink node. Massa and Avesani [105] developed MoleTrust, which predicts one's trust towards another by walking through the social network and propagating trust values along trust edges.

7.2.2 Credit Over Risk Equals Trust

Local trust models use personalized trust measures to accommodate diverse subjective views on the same node, while global models use an objective reputation to approximate how much a social network as a whole trusts each node. While existing models are primarily based on either global or local trust inference, in most social networks, both the subjective and objective properties of trust should be considered in order to infer personalized and more accurate trustworthiness of each node. The novelty of the CoreTrust model is to propagate trust in a distributed manner within a social network in the form of credit flow. Suppose A, B, and C are three nodes in a social network. A would trust B if credit flowed from A to B, e.g., when A gave credit to B through good ratings or positive comments. If C's risk factor were lower than B's, e.g., because C is an expert, whereas B is a novice in a particular field, A would probably trust C more than B. Therefore trust is decided by the credit flow and the risk disparity between B and C. A

node would be more trustworthy if more credit flowed into that node or that node had a lower risk factor.

Figure 7.1 compares the differences among global, local, and credit-flow-based trust models. Figure 7.1(a) describes the acquisition of objective metrics using an algorithm similar to PageRank. A node's PageRank value, shown as a number within the node, can be obtained through voting, where a vote is a trust link from any other node to this node, e.g., PR(A)=(1-*d*)+*d*×((PR(B)/N(B)+(PR(C)/N(C)), where *d* is the damping factor (0.85 in this example), and N(B) and N(C) are the numbers of outbound links from B and C respectively. Figure 7.1(b) illustrates the distributed propagation of subjective metrics using a strategy where the trust value between a pair of indirectly connected nodes can be calculated from existing trust values between directly connected nodes. In this example, the trust value from node A to node D, shown as a dashed edge, is derived from existing metrics: 0.4×0.6+0.1×0.3=0.27.

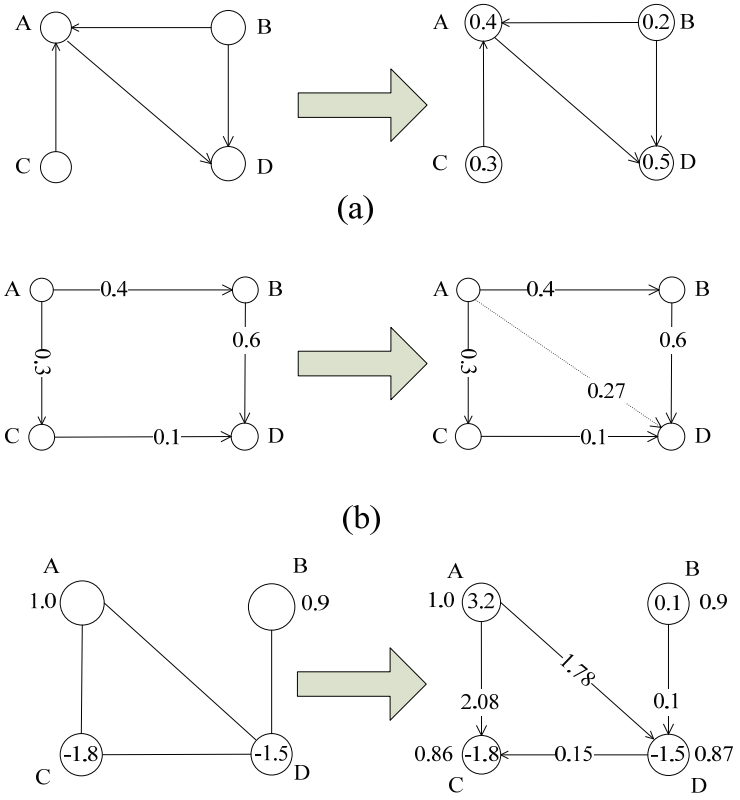

Figure 7.1 (a) global trust model; (b) local trust model; (c) credit-flow-based trust model

Figure 7.1(c) depicts the distributed inference of trust values using credit flow. A node's accumulated credit, e.g. -1.8 and -1.5 for nodes C and D respectively, is derived from the evaluations it received from other nodes. A node's risk factor, e.g. 1.0 for node A and 0.9 for node B (relative to A), in other words, compared to node A, node B has a lower risk factor, is derived from the expertise level of that node against that of another node it is compared to. Assume that node B's credit is 0.1 and credit can only flow from node A to other nodes. With the CoreTrust model, we can infer that credit sourced from A is 3.2 (1.8+1.5-0.1), with 1.65 and 1.55 flowing into C and D respectively and that C's and D's risk factors (relative to A) are 0.86 and 0.87 respectively. We can further infer that trust

values from A to C and D are 2.08 ((1.65+0.15×1.55/(1.55+0.1))/0.86) and 1.78 (1.55/0.87) respectively. Details about how the credit, risk, and trust values are derived from the credit flow model will be given in subsequent sections, but it is worth pointing out that the trust values are inferred from both the objective ground, e.g. node C's and D's credit accumulated from their interactions with other nodes in the whole network, and the subjective measures, e.g. node C's and D's risk calculated from their expertise disparity against A.

The credit-flow-based trust model is inspired by the power flow in an electrical grid. Our hypothesis is that the credit flow in a social network is fundamentally similar to the power flow in an electrical grid. If we map users, credit, risk, and trust in a social network to devices, power, voltage, and current respectively in an electrical grid, we can observe their commonalities. For example, some users give credit to a social network; while others take credit from it as if some devices generate power; while others consume power in an electrical grid. Trust flows from users of high risk to users of low risk as if current flows from devices of high voltage to devices of low voltage. The relationship among credit, risk, and trust in a social network is credit = risk × trust (more credit or low risk infers high trust), which is analogical to the power = voltage × current relationship in an electrical grid.

To test this hypothesis, we will infer node trust relations in a social network using the physical and mathematical principles in power flow study [139] – an important tool for analyzing the voltages, power flows, and currents in an electrical system under steady-

state conditions. We will present the credit-flow-based trust model, and some background of power flow study in electrical systems is introduced in Appendix G.

7.3 The Credit-flow-based Trust Model

The credit-flow-based trust model CoreTrust was inspired by the analogy between the distribution of credit in a social network and the propagation of power in an electrical grid. We first present the constituents of our model, which were motivated by the elements in power flow study. We then describe the credit balance equations formulated from the model and finally discuss distributed inference of trust values using the credit balance equations.

7.3.1 The CoreTrust Model

Most online communities, such as Weblogs Blogger and Blogosphere, shopping sites eBay and Amazon, social media sites YouTube and Digg, review sites Epinions and Slashdot, peer-to-peer networks eDonkey and BitTorrent, and our epistemology-based social search community, allow members to rate the content generated by others. Global trust models have used the rating information to infer each node's universal reputation, but our model utilizes such information to derive credit and further to trace the credit flow in order to infer personalized and accurate trust values.

For the sake of clarity without losing generality, we take Epinions[16], a popular e-commerce site where users can write reviews on various items such as cars, books, or music, and rate the items or reviews, as a reference social network to describe the CoreTrust model and to evaluate the performance of the model against existing models.

The CoreTrust model is represented by the key elements of credit, risk, bias, and trust, corresponding to power, voltage, phase angle, and current respectively in power flow study.

Credit: represents the confidence a user has in all other users. In Epinions, it can be derived from each user's ratings given to other users, that is the credit brought into the network by the user, or each user's ratings received from others, that is the credit taken out by the user. In a social network, a user's confidence may vary in different context, e.g., different categories in Epinions. Given a specific category, such as music, a user's credit in this category is real credit C (analogous to the real power P in an electrical network), and the user's credit in other categories is reactive credit D (analogous to the reactive power Q in the electrical network). To simplify the model, we only consider the net credit brought into the network, i.e. the credit brought in by a user minus the credit taken out by the same user. This is reasonable as a balance is kept for the total credit in the network. The formulae for calculating C and D are defined as:

$$C_u = \sum_{rev \in Ca} RA_u^{rev} - \sum_{rec \in Ca} RA_u^{rec} \qquad (8)$$

[16] http://www.epinions.com

where *rev* and *rec* are reviews in a category *Ca*, RA_u^{rev} is the rating given by user *u* to the review *rev*, and RA_u^{rec} is the rating received by user *u* for her/his review *rec*.

$$D_u = \sum_{rev \notin Ca} RA_u^{rev} - \sum_{rec \notin Ca} RA_u^{rec} \qquad (9)$$

where *rev* and *rec* are reviews in categories other than *Ca*.

Risk: represents the possibility of incurring loss or failure when believing a user. Risk of believing someone is intuitively related to the expertise, i.e. the level of one's expert knowledge or skill in a particular field, of the person to be believed. Generally speaking, one with a high level of expertise would have a low risk of trustworthiness. Suppose the expertise *E* is the degree of competency to provide accurate ratings and exhibit high activities in an online community such as Epinions [88], a user's risk factor *R* can be derived from her/his reviews on items in a category.

First, if an item *i* received ratings from *N* users, each providing R_u^i for it, the average rating of this item is:

$$RA^i = \frac{1}{N} \sum_{u=1}^{N} RA_u^i$$

If user *u* provided a rating for item *i*, we can measure her/his expertise level by comparing her/his rating to the average rating from all other users, i.e. excluding user *u*:

$$RA_{\neg u}^i = \frac{1}{N-1} \sum_{u=1}^{N} RA_u^i - RA_u^i$$

A small difference between the two ratings suggests a high level of expertise for user *u*.

Next, the following formula defines the expertise of rating one item (item *i*):

$E_u^i = 1 - \frac{\left| RA_u^i - RA_{\neg u}^i \right|}{RA_{Max}^i}$, where R_{Max}^i is the maximum rating scale.

Finally, if user *u* provides ratings for *M* items in a category, the risk of believing *u* is defined by the accumulated expertise of rating the *M* items:

$$R_u \propto \frac{1}{E_u} = M \Big/ \sum_{i=1}^{M} E_u^i = M \Big/ \left(1 - \frac{\left| RA_u^i - RA_{\neg u}^i \right|}{RA_{Max}^i} \right) \qquad (10)$$

Bias: represents a user's preference. For example, in the music category, if one is interested in classic music and another is interested in pop music, the bias between the two users would be 90 degrees. Bias β is defined by the following formula:

$$\beta = \arccos\left(\frac{\mathbf{W_a} \cdot \mathbf{W_b}}{\|\mathbf{W_a}\| \|\mathbf{W_b}\|} \right) = \arccos\left(\frac{\sum_{i=1}^{n} W_u^i \times W_b^i}{\sqrt{\sum_{i=1}^{n} (W_a^i)^2 \times \sum_{i=1}^{n} (W_b^i)^2}} \right) (11)$$

where $\mathbf{W_a}$, $\mathbf{W_b}$ are items reviewed by user *a* and user *b* respectively, *n* is the total number of items reviewed by *a* or *b*, W_u^i=1 (if user *u* reviewed item *i*) or 0 (if user *u* did not review item *i*). In particular, β=0 if *a* or *b* hasn't reviewed any item.

Trust: represents how much a user can believe another user in a particular field. Similar as a current in an electrical grid, trust is also directional in a social network. For example,

Alice may not trust Bob in the same way Bob trusts Alice. In the CoreTrust model, the trust value from the trusting to the trusted is inferred from all the relevant trust flows.

7.3.2 Credit Balance Equations

Users in an online community can be classified as source user, appraisal user, or beneficiary user, corresponding to the three types of buses in an electrical network: slack bus, generator bus, and load bus.

Source user: a user whose trust values towards other users need to be inferred. A source user's risk and bias are set to 1 and 0 respectively. A source user's credit is yet to determine as it is unknown in the beginning how much credit the user would bring into the network, but it is assumed that user can adjust real credit C and reactive credit D in order to keep a good balance.

Appraisal user: a user who evaluates other users by giving credit to them. It is assumed that an appraisal user's risk is known a priori. In the initial status, all credit in the network is brought in by the appraisal users. When inferring trust values for the source user, we suppose all credit is brought in by the source user to achieve the credit balance in the network. Therefore there is no need to calculate an appraisal user's credit and it is set to a minimal value, i.e.: 0.1.

Beneficiary user: a user who is evaluated by other users through receiving credit from them. A beneficiary user's credit is negative as they take credit from instead of contributing credit to the network.

It is worth clarifying two points. First, users who have not been evaluated by others and who have evaluated others but have shown no expertise are excluded from our model as their credit could not be measured or verified. Second, the definitions of real and reactive credit refer to net credit, i.e. the credit brought in by a user minus the credit taken out by the same user. Therefore, an appraisal user would have positive credit, whereas a beneficiary user would have negative credit.

Inspired by the power balance equations, we can formulate the credit balance equations:

$$C_i - jD_i = R_i^* \sum_{k=1}^{N} Y_{ik} R_k \qquad (12)$$

where N is the number of users in a social network, C_i is the net real credit given to user i, D_i is the net reactive credit given to user i, R_k is user k's risk factor, and Y_{ik} is the total admittance between user i and k, which can be found from the user admittance matrix $\boldsymbol{Y}_{user}$ of the network. Give a source user among all the users, the admittance matrix is built according to the following two rules: (1) the admittance of elements connected between user k and the source user is added to entry (k, k) of the admittance matrix, and (2) the admittance of elements connected between users i and k is added to entries (i, i) and (k, k) of the admittance matrix, while the negative admittance is added to entries (i, k) and (k, i) of the admittance matrix.

The rationale behind the credit balance equations is the perfect balance between the credit brought into the network and the credit taken out of the network. That is all credit brought into the network is properly distributed to the nodes that consume them. Credit distribution in a social network follows the rules similar to those for power distribution in

an electrical network. On the one hand, if credit is fixed, lower risk infers more trust, which is similar to an electrical network, where lower voltage leads to stronger current if the power is fixed. On the other hand, if risk is fixed, more credit infers more trust, which is similar to an electrical network, where high power leads to stronger current if the voltage is fixed. For example, we would trust an unfamiliar person if they were acclaimed by a well-known expert (i.e., the risk of believing this expert is low) or by lots of people (i.e., the credit of this person is high).

7.3.3 Distributed Trust Inference

With reference to the steps and equations described in power flow study, we now reformulate the credit flow equations:

$$C_i = \sum_{k=1}^{N} |R_i||R_k|(G_{ik}\cos\beta_{ik} + B_{ik}\sin\beta_{ik}) \qquad (13)$$

$$D_i = \sum_{k=1}^{N} |R_i||R_k|(G_{ik}\sin\beta_{ik} - B_{ik}\cos\beta_{ik}) \qquad (14)$$

where G_{ik} and B_{ik} are the real and imaginary parts of the admittance matrix element Y_{ik} respectively, and β_{ik} is the bias of user i towards user k.

We then use the Newton-Raphson method, which is well known for its convergence characteristics and speed, to derive each beneficiary user's risk and all credit flows within a social network. As a standard Newton-Raphson method can be implemented with hardware components for solving systems of linear equations, it could be a simple and reliable way to estimate the risk and credit in our model. To do so, we first determine the

normalized trust value between every pair of directly connected users using the following formula: $T_{ij} = \hat{I}_{ij} \propto \frac{R_i - R_j}{r}$

where r is user i's resistance factor against user j. Resistance is a psychological factor describing one's instinctive distrust towards a stranger. The further away user A is from user B in a social network, the higher resistance (or the weaker trust) A would place on B because trust can be diluted over a long propagation path [79]. For example, if a propagation path consists of 10 nodes and the trust value from one node towards the next in the path is all 0.9, the last node will only receive the trust value of $0.9^{10} \approx 0.35$, and for this reason, some trust models such as *MoleTrust* [105] limit the length of a propagation path using a maximum depth in search.

The resistance factor is an analogy to the impedance factor associated with a transmission line in an electrical network, which causes power loss in transmission. However, it is non-trivial to model resistance in a social network because many factors need to be considered and they are generally difficult to estimate accurately. For the sake of simplicity, the resistance factor between every pair of users is set to a constant and small value r=0.05+j0.2 in the above formula, where 0.05 and 0.2 are empirical coefficients drawn from experimental data and j is the imaginary unit for irrelevant credit. Credit dilution is related to the risk difference in that a higher risk suggests more diluted credit. The two empirical coefficients also suggest that irrelevant credit is diluted faster than relevant credit.

According to the Kirchhoff's circuit law, the credit flows out of a node equals to the credit flows into that node, and we can infer the trust value between each pair of indirectly connected users as such. As shown in Figure 7.2, node *i* is directly connected to node *k* and node *k* is directly connected to node *j*. If the credit flows from *i* to *k* and from *k* to *j* are C_{ik} and C_{kj} respectively, and the total credit flowing into *k* is C_k^{all}, the credit flowing from *i* to *j* is: $C_{ij} = C_{ik} \cdot \frac{C_{kj}}{C_k^{all} - C_k}$

and the trust value from *i* to *j* is: $T_{ij} = T_{ik} \cdot \frac{T_{kj}}{T_k^{all} - T_k}$

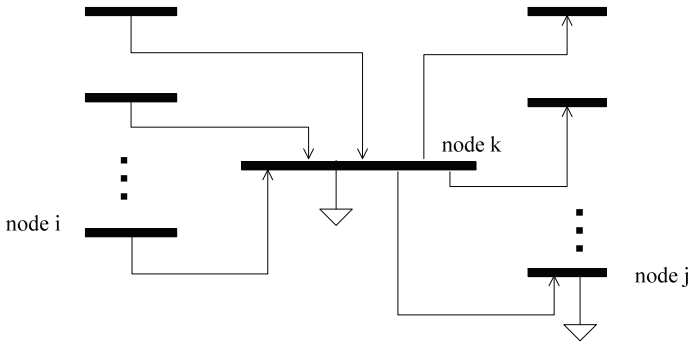

Figure 7.2 An example of trust propagation via one node k.

If node *i* has *M* directly connected nodes and node *j* has *N* directly connected nodes, the trust value from *i* to *j* is:

$$T_{ij} = \sum_{k=1}^{M} T_{ik} \cdot \frac{\sum_{k=1}^{N} T_{kj}}{\sum_{k=1}^{M+N} T_k^{all} - \sum_{k=1}^{M+N} T_k} \quad (15)$$

For example, a simple social network is shown in Figure 7.3 and we want to infer Alice's trust values towards other users in the network, i.e., Alice is the source user. Suppose we have already calculated Bob's, Carol's, Dave's and Eve's credits to be 0.1, -3.2, -1.6, and

-2.0 respectively. We can then derive that the real credit brought into the network by Alice is 6.9. If Bob's risk is 0.95 (relative to the source user's risk 1.0), we can also derive Carol's, Dave's and Eve's risks to be 0.59, 0.62, and 0.38 respectively. Finally we can infer Alice's trust value towards Carol, Dave and Eve to be 0.51, 0.26, and 0.23 respectively.

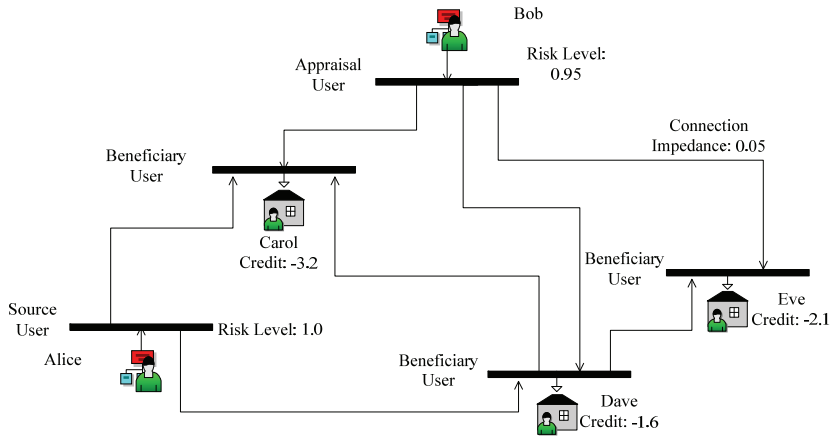

Figure 7.3 Distributed Trust inference in a simple social network

In contrast, global trust models compute each user's universal reputation without considering a source user's diverse preference and expertise. Local trust models compute personalized trust values between indirectly connected users purely from the trust values between directly connected users without considering each individual user's diverse expertise and credit derived from their real interactions.

In summary, distributed trust inference using the credit balance equations formulated from the credit-flow-based trust model is done through the following steps:

1) *Calculate the real and reactive credit of each user in a social network based on formulae (1) and (2).*

2) *Classify users into the source user, the appraisal users, and the beneficiary users according to their credit and expertise. Calculate the expertise level and bias of each appraisal user based on formulae (3) and (4), relative to the source user whose expertise level and bias are set to 1 and 0 respectively.*

3) *Build the user admittance matrix* $\mathbf{Y}_{user}$ *based on the topological structure of the social network and form the credit balance equations (5).*

4) *Use the Newton-Raphson method to solve Equations (6) & (7) and the expertise levels of all beneficiary users, and infer the trust values from the source user to all directly connected users.*

5) *Infer the trust values from the source user to all other users through distributed trust propagation based on formula (8).*

6) *Choose a different source user and perform the distributed trust inference again (iterating from step 2) until all users have been considered.*

Figure 7.4 The distributed trust inference algorithm

7.4 Evaluation

We have conducted a set of experiments to compare the quality of trust inference between the CoreTrust model and existing models with objective and subjective trust metrics in order to answer the following two questions:

1. How does the credit-flow-based trust metrics compare to the objective trust metrics in terms of accurately identifying a user's trustworthy peers in a social network based on the reputation ranking of all other users in the network?

2. How does the credit-flow-based trust metrics compare to the subjective trust metrics in terms of predicting a user's trust towards others in a social network based on the distributed trust propagation between users in the network?

7.4.1 The Epinions Dataset

The Epinions dataset was used as the trust data in our experiments. This dataset, consisting of various types of user interactions and rated objects in multiple categories, combines explicit user ratings and trust sets with social networking. Similar to our *Baijia* prototype system of the EPISOSE framework, Epinions allows users to write their own reviews or rate reviews from others by assigning a helpfulness rating from 1 (not helpful) to 5 (most helpful). To encourage high quality reviews, the Income Share program is adopted to pay users according to the ratings of their reviews. Consequently trust management becomes a critical issue and Epinions allows a user to specify how much she/he trusts others and uses the resulting web of trust to sort the product reviews for the user.

We developed a program using the techniques contributed by Richardson [134] to retrieve the trust relationships in Epinions by crawling the website. The dataset contains 28,377 users who contributed a total of 258,122 reviews on different categories of products, which received a total of 1,784,182 ratings. Due to the computational

complexity, we only chose the "Video & DVD" category in our experiments as it has more reviews per product than other categories. Only users who contributed at least one review or rated at least one review are included in the dataset. A record in the dataset is either <review_id, contributor_id> where review_id is the ID of a review on a subject, e.g., a movie in the category and contributor_id is the ID of the user who contributed the review or <review_id, rating, evaluator_id> where rating is in the scale from 1 to 5 and evaluator_id is the ID of the user who gave the rating.

We then extracted the trust relationships among these users, which formed a web of trust graph consisting of 6,847 vertices and 77,965 edges. The graph is stored as records of <trusting_id, trusted_id> pair, where trusting_id is the ID of the user making the trust statement, trusted_id is the ID of the user on which the trust statement was made, and the trust relationship is a directional edge from the trusting_id vertex to the trusted_id vertex. In our web of trust graph, 2,879 vertices had at least one outbound and inbound edges and 4,576 vertices had at least one inbound edge.

It is worth pointing out that Epinions users can also keep a set of distrust relationships (or blacklist) by specifying users they do not trust at all, but unlike Guha's work [59], we did not exploit distrust relationships and instead only focused on the trust relationships in our experiments primarily because the amount of distrust relationships are rather small, for example in the dataset from Victor et al. [169], about 85% of the statements are labeled as trust. Nonetheless, we will investigate whether distrust relationships can be used to improve our trust model in the future.

The dataset confirmed the similarity between a social network and an electrical grid that they both follow the power law distribution [124] and are both small-world networks [176]. For example, we found that most users contributed very few reviews or ratings and that very few users contributed extremely many reviews or ratings. Similarly, we found that most users trusted very few others and that very few users trusted extremely many others.

7.4.2 Comparing with Global Trust Models

We first compare the prediction accuracy between the credit-flow-based model and global trust models using well-defined metrics such as *Precision*, *Recall*, and *F1*. Given a prediction result as a ranking of the topmost trustworthy users in a network, we define:

$$precision = \frac{\{\text{real trusted users}\} \bigcap \{\text{predicted topmost trustworthy users}\}}{\{\text{predicted topmost trustworthy users}\}}$$

$$recall = \frac{\{\text{real trusted users}\} \bigcap \{\text{predicted topmost trustworthy users}\}}{\{\text{real trusted users}\}}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{recall + precision}$$

It is worth clarifying that the list of users who have been given trust statements by a user is much smaller than the list of users who are really trusted by this user, because the number of explicit trust statements in Epinions is rather small. Therefore it is difficult to exactly measure the accuracy of trust prediction because we cannot simply assert whether a user trusts or does not trust other users without referring to explicit trust statements. To

overcome this difficulty, we treat the user trust statements as the ground truth and compare them with the objective reputation ranking of all users in the network by adjusting the length of topmost ranking according to each user's preference. For example, if a user made trust statements about 5 users, we can measure the prediction accuracy based on the predicted top 10 most trustworthy users.

In this experiment, we compare our CoreTrust method with the following two global trust methods:

- The *Average* method, which ranks all users in such a way that is similar to the simplified eBay model. It evaluates each user according to the ratings the user has received from others.
- The *EigenTrust* method [81], which assigns a global trust rating to each user using an algorithm similar to PageRank. It aggregates users' trust information through performing a calculation approaching the eigenvector of the trust matrix over the users.

Because these methods provide a unified objective ranking of reputation for all users, but our model provides personalized ranking of trustworthiness for every individual user, we first derived unified and personalized trust values using the three methods respectively, then calculated the *Precision*, *Recall*, and *F1* metrics for each user, and finally took the average of the three metrics for the whole network.

Table 7.1 shows the prediction accuracy of the three methods, where the *EigenTrust* method performs slightly better than the *Average* method, but the *CoreTrust* method

significantly outperforms the other two, especially for the *F1* metrics, where both the precision and the recall are taken into account. The *Average* and *EigenTrust* methods assign a universal trustworthiness score to each user and every other user in the network has to trust it in a unanimous way that is decided by the score. The reality is that in the Epinions community, individual users have different preference and expertise and may be trusted by different users in different ways. The *CoreTrust* model has taken into consideration that fact and allows multiple trustworthiness scores to be assigned to each user, one for every source user. Therefore, the average *Precision*, *Recall*, and *F1* metrics for the *CoreTrust* model are much better than those for the *Average* and *EigenTrust* methods.

Table 7.1 Prediction accuracy

Methods	Precision	Recall	F1
Average	0.47	0.61	0.53
EigenTrust	0.53	0.72	0.61
CoreTrust	0.79	0.82	0.81

7.4.3 Comparing with Local Trust Models

We then want to compare the trust propagation accuracy between the credit-flow-based trust model and local trust models based on subjective trust metrics. In local trust models, explicit trust values are specified for directly connected users, while trust values for indirectly connected users need to be inferred through distributed trust propagation. In our experiments, we use two trust networks: one is the original web of trust collected

from Epinions and the other is the web of credit constructed from the credit-flow-based trust model. We use the subjective metrics to propagate trust over both networks and compare their results.

This evaluation is a kind of leave-one-out cross-validation, where we are given a network with all (but one) of the trust values between nodes visible, and we need to predict this single suppressed value. The original web of trust in Epinions is a collection of binary ratings, i.e., trust (1) or non-trust (0). We consider two commonly used metrics for evaluating binary classifiers: the classification *accuracy* of predicting correct ratings and *AUC*, which represents the area under the receiver operating characteristic curve (ROC curve). If we define the number of trust users and non-trust users as *P* positive instances and *N* negative instances respectively, a *true positive* (*TP*) occurs when both the prediction outcome and the actual value are trust and a *true negative* (*TN*) occurs when the prediction outcome and the actual value are both non-trust. Therefore, the *accuracy* is defined as:

$$ACC = (TP + TN)/(P + N)$$

The *AUC* value is calculated as the integral of ROC curve, which equals to the probability that a user who should be trusted is correctly predicted as trust rather than non-trust. Therefore it outputs probabilities instead of binary decisions.

In this experiment, we compare our CoreTrust method with the following two local trust methods:

- The *Direct* method, which propagates trust by direct propagation. For example, if Alice trusts Bob and Bob trusts Carol, then trust propagates from Alice to Carol.
- The *MoleTrust* method [105], which propagates trust by combining trust ratings across all paths from a source vertex to a destination vertex in a graph. Paths are searched in a typical breadth first search fashion (maximum depth is set to 2 in our experiment).

Because a trust value is binary in the web of trust and non-binary in the web of credit, a threshold is used in the web of credit to determine between trust and non-trust. In the *Direct* method, trust is propagated as long as the inferred trust value in the web of credit is greater than the threshold. In the *MoleTrust* method, paths with the aggregated trust values above a threshold are selected for trust propagation. A tradeoff can be made between *TP* and *TN* by adjusting the threshold. We used a consistent predefined threshold to calculate the *accuracy* and *AUC* metrics and results are shown in Figure 7.5 and Figure 7.6 respectively.

The *Direct* and *MoleTrust* methods were tested for prediction accuracy on both the Epionions web of trust and the web of credit constructed using credit-flow-based model. Figure 7.5 shows the two methods' prediction accuracy on the web of credit is clearly better than that on the web of trust, confirming that the credit-flow-based model can construct more accurate trust connectivity from an online community with users' rating data. This may attribute to the sparse nature of the Epinions web of trust because the number of explicit trust statements made by each user in Epinions is rather small. In

contrast, the web of credit is much denser because the credit-flow-based model can derive significant more trust relationships from the rating data.

Figure 7.5 Accuracy evaluated on the web of trust and the web of credit

More importantly, comparing the *Direct*, *MoleTrust,* and *CoreTrust* methods tested on the web of credit, the *CoreTrust* method clearly outperforms the other two. This is because the *CoreTrust* method considers both generally agreed reputation and individual users' preference and expertise, while the other two methods only consider subjective trust statements.

The performance comparison is further confirmed by the results for *AUC* metrics in Figure 7.6, which are nearly the same as those for the accuracy metrics in Figure 7.5.

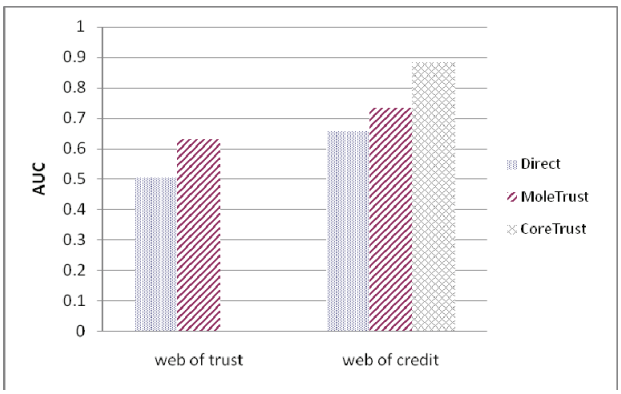

Figure 7.6 AUC evaluated on the web of trust and the web of credit

7.5 Summary

In our epistemology-based social search solution, social network services are in place to support epistemology sharing between prosumers in the social search community. Trust management in social networks is imperative as the epistemology quality is closely related to the trustworthiness of the user who contributes or refines the epistemology.

Inspired by the physical and mathematical properties and the power flow study in electrical grids, we proposed the credit flow approach to modeling and inferring trust relations in social networks. This approach can construct a web of credit from the interactions in a social network, such as the contribution and evaluation of epistemologies, and infer trust values by making use of both generally agreed reliability and subjective individuality in the network. The experimental results on a real-world dataset show that the approach can accurately construct a web of credit and infer more accurate trust values than objective and subjective trust models do.

The credit-flow-based trust model has been applied to our epistemology-based social search system. It is possible to reduce the computational complexity involved in our approach if more efficient algorithms for solving the credit balance equation (e.g., Fuzzy Logic Control) are adopted in the future.

Chapter 8

Non-monetary Incentive Mechanism

In social search systems, an important consideration is in regard to the incentives of users to share their knowledge with others. An effective incentive mechanism to encourage users to contribute and refine epistemology is necessary in the epistemology service component of the EPISOSE framework. As a user's experience or knowledge that can fulfill others' information requirements is of great worth, it is natural that the owner should not share it freely and would like to ask remunerate for it. Hence we propose a novel incentive mechanism for social search systems based on the epistemology trading, where users are encouraged to contribute their knowledge to a social community and can trade it for interested knowledge of others. Furthermore, we propose the online silk road to model the epistemology trading, and develop services and algorithms to help users acquire knowledge while maximizing the social welfare in the social search community. Experimental results demonstrate the feasibility and effectiveness of the epistemology-trading-based incentive mechanism.

8.1 Introduction

A great deal of users who will share their intimate knowledge with others in seeking for information is one of the necessary conditions of a successful social search system. However, the incentive to contribute the high-quality knowledge is a critical issue and

has not been researched in depth before. A commonly held belief is that users are willing to share information for free, and since free information is everywhere, it is impossible and meaningless to ask for payment when publishing information. The plenty of UGC available on numerous social media websites is a good proof. However, we will see that information is not necessarily free to publish or access and it is unreasonable that a user would like to give away information which is useful to others with specific requirements. Therefore an incentive mechanism which can encourage users to contribute knowledge is essential, and more importantly, the knowledge is authentic as the owners will be responsible for the explanation and maintenance of the knowledge.

"Information wants to be free" has been the mantra of the Internet at the very start, and people have become used to feasting on online freebies of all sorts: news, stories, and videos. However, useful information doesn't emerge of itself and has to be gathered, processed and checked by professional or amateur reporters. It was once believed that websites would make truckloads of cash from online advertising. Unfortunately, the revenues from online advertising turns out to be not enough to cover the cost of even a tiny news operation, especially after the financial crisis. As an example, Rupert Murdoch's News Corporation has erected a "pay wall" around its UK newspaper titles including *The Times* and *The Sunday Times*, in which readers will be charged to access online articles. SEs such as Google are also blocked and banned from linking with the content to further fend off freeloaders. More recently, the *New York Times* turned a profit since it launched a paywall on its website in March and now has 324,000 paid digital subscribers.

Although the online payment strategies that intent to end the free Internet era will continue to be controversial before the new bubble burst, many online media websites actually are seeking for revenues from other sources, such as online games. More importantly, for most of the users who publish information on those social medial websites, there are no clear revenue models for them to get benefit from their content. Therefore they just generate the content voluntarily and unceremoniously. Much of such content not only makes the information overload problem on the Internet more serious, but also probably contains imprecise information that will mislead others seeking for related content.

Therefore knowledge and useful information, similar to common goods, can have monetary value because skilled labor is required to create them. We can view and design the incentive of contributing and acquiring knowledge based on the study of economics. On the one hand, although information is everywhere, information contains valuable knowledge to others is rare and much information published by the majority of users is not or only partially to others' requirements. On the other hand, the example of the *New York Times* paywall indicates that users are willing to pay for the information they really want. Since the Internet has provided people from around the world not only a source to acquire knowledge and a place to publish information, but also an environment to communicate with each other, the Internet-based knowledge markets are emerging where users having original knowledge and other users willing to pay for it can be connected. The remunerative incentives rather than moral or coercive incentives are the main form of incentives employed in various markets, and can be employed in knowledge markets too.

The most popular remunerative incentive mechanisms in knowledge markets are points-based incentive programs, which is a type of program where participants can collect and redeem points for rewards. The virtual currency is the foundation in the markets. The idea is that when users contribute any content, they will get some electronic money or points, which can be used to exchange for something in real-world such as goods and phone charges, or to purchase content contributed by others.

However, the virtual currency has some drawbacks and one of the inescapable problems is inflation. Since users of a social media website could seize a lot of virtual currency through collusions or sockpuppets, the amount of virtual currency is much greater than the worth of real effective information on the website. The price of information will increase continuously and lead to inequalities in wealth. It is to newcomers' disadvantage and will harm the stability of the society. Actually, as "inflation is always and everywhere a monetary phenomenon", even the economic systems in real world with the central bank's control of the money supply could not prevent it.

In this chapter, we propose a novel incentive mechanism for social search systems based on the epistemology trading, where prosumers can trade the knowledge they possess for the knowledge they require by describing what they are supplying and demanding in their epistemologies. Optimal trades can be identified to maximize the social welfare of the community where most prosumers can acquire more relevant knowledge they required. This approach escapes the traps of the modern economics, and is inspired by the inter-regional trading activities on the ancient *Silk Road*. Along the *Silk Road* routes, China got precious stones of India, India got gold and silver of the Roman Empire, and the Roman

Empire got silk and porcelain of China. All these trades were performed based on the item-to-item bartering, as the foreign currency market was completely not existent.

Barter markets still exist in many countries, but only to a very limited extent, comparing to monetary systems in real world. However, our approach has addressed the following limitations of bartering, which are acknowledged as the main reasons for replacing barter with currency. Therefore the non-monetary knowledge trading is a feasible mechanism for online information sharing and acquisition.

First, the absence of common measure of value in a barter economy makes it difficult to exchange different kinds of goods. However, it is insignificant in knowledge trading since naturally knowledge cannot be measured accurately. The knowledge worth 10 dollars is not necessarily better than the knowledge worth 5 dollars. Therefore we leave the measure for users: they can exchange their knowledge with others' they think is valuable to them, and they do not need to consider the exchange rates.

Second, difficulty in storing wealth is an issue when the time and place don't coincide in barter of perishable goods. This is not a problem for knowledge trading in the Internet era, as users can trade with anyone from anywhere in anytime, and they can store their knowledge in our system for the future trading.

Third, need for presence of double coincidence of wants causes the low efficient of bartering because a lot of time is wasted in searching for the transaction target. In our approach, we propose an efficient algorithm to automatically find the transaction target for users based on the cycle formulation in a graph. The experiments based on real world

dataset has invalidated the effectiveness of the algorithm and proved the practicality of the incentive mechanism based on knowledge trading.

8.2 Related Work

8.2.1 Internet-based Knowledge Markets

In the discipline of information management studies, knowledge sharing has been related with social exchange theory. That is, the knowledge movement in the information society is powered by market forces. The concept of knowledge markets, proposed by Davenport and Prusak [40], depicts a mechanism for supporting and facilitating the sharing or exchange of knowledge among users, where organizational actors are defined as knowledge buyers or sellers within a marketplace. Unlike users in the markets for more tangible goods, knowledge buyers are individuals trying to resolve an issue with complexity that precludes an easy answer, and knowledge sellers are people with a reputation for having substantial knowledge about a subject or process. As both buyers and sellers believe they can benefit from sharing the intellectual capital, some trading exists in knowledge markets [110].

The effective operation of knowledge markets has brought the knowledge flowing inside the knowledge net and realized the sharing of knowledge [71]. To improve business processes, knowledge management can be used to increase productivity and reduce new product development times. For example, Müller et al. [8] analyzed different quality management methods for marketplaces of knowledge, and discuss the commonalities and differences between traditional knowledge management systems and knowledge markets.

Compared to traditional goods, knowledge exchange is more convenient on the Web. In addition, computer-based actors such as software agents that enact as buyers and sellers can be involved in Internet-based knowledge markets. The pervasiveness of the Web has started to shift existing knowledge markets into the Internet [154].

In recent years, a number of studies involve the implications and impact of online knowledge markets from the inspirations of Web-based question answering communities. Most QA websites, such as *Yahoo! Answers*, *Windows Live QnA*, and Naver's *Knowledge-iN*, use free knowledge exchange models, which only offer an increase in reputation as payment for researchers. *Google Answers* was a service that allowed users to offer bounties to expert researchers for answering their questions. The *Google Answers* site was closed in 2006 but former *Google Answers* researchers launched the paid QA/research site *Uclue*[17] later, where customers will pay via PayPal when they post a question. *ChaCha*[18] offers subsidized knowledge markets where the question asker is given free answers in an interactive and conversational format, while the guides are paid to generate answers. Zhang and Jasimuddin [199] proposed a mathematical model for investigating the working mechanism behind an online knowledge market, i.e., the pricing strategies of users and the company who maintains the market. Nevertheless, no work has been done on the prerequisites for the market behavior as the incentive of knowledge exchange.

---

[17] http://www. uclue.com
[18] http://www. chacha.com

8.2.2 Virtual Currency

Virtual currency has been used in online knowledge markets for knowledge trading. For example, the *Experts-Exchange* [19] market, which has pioneered the information technology professional network marketplace since 1996, aims to provide specific solutions to specific problems using a virtual currency. It provides a marketplace where buyers may offer payment to have their problems resolved. *Mahalo Answers*[20] made the QA model work financially by launching a paid Answers service, which is an extension of the people-powered SE Mahalo.com. *Mahalo Answers* users can provide a monetary reward in the form of its proprietary currency - Mahalo Dollars. In fact, virtual currency has been widely used to purchase virtual goods within a variety of online communities including online gaming sites, virtual worlds and social networking websites.

Wang and Mainwaring [174] defined virtual currency as private currency intended for online use. Their investigation has shown that the money aspect of user experience in online games closely interrelates with other important aspects such as fairness and fun. Yang et al. [193] described the symbolic meaning of virtual currency to the interpersonal relationship in an online bulletin system. Yamaguchi [191] has argued that there are no qualitative differences between virtual and real currencies. Wang et al. [172] presented a quantitative study by examining the virtual value generated from the time spent on playing games. Most of these researches are economics-oriented and did not focus on the technical point of views on virtual currency exchange.

---

[19] http://www. experts-exchange.com
[20] http:// www.mahalo.com/answers

Irwin et al. [72] suggested that economic problems in real world such as inflation and deflation need to be carefully considered in virtual currency models. Because it is easy to raise prices on virtual goods or increase money supply, these models often suffer hyperinflation and other economic issues [111]. Virtual currency not only plays the role of a facilitator of the transactions in virtual worlds but also as a link to connect the virtual economy to the real world economy [153]. Guo et al. [61] proposed methods of realizing virtual wealth in both virtual and real worlds based on the labor theory of value, exchangeable value, and utility theory. In real-money trading (RMT) markets, virtual currency can be officially exchanged with real money in some social virtual worlds such as the 3D-animated world of *Second Life*. To attract existing players, one of the core strategies of online games is to add the social hierarchy for avatars. However, this strategy stimulates players to shortcut game-play through RMT. Therefore it creates the demand for gold farming which exacerbates the inflation in virtual worlds and reduces a game's lifecycle [33].

8.3 Epistemology Trading Silk Road

Considering users of a social search community, such as prosumers in *Baijia* system, each of them is seeking some knowledge on one subject while holding some knowledge on another subject. Our incentive mechanism is to encourage prosumers contribute the knowledge they possess and enable trades among prosumers so that all participants will benefit from contributing knowledge. These trades build up cycles of prosumers (the knowledge trading silk road), with each prosumer receiving the knowledge of the next prosumer in the cycle. For example, prosumer A receives the knowledge of computers

from prosumer B, prosumer B receives the knowledge of medicine from prosumer C, and prosumer C receives the knowledge of history form prosumer A. The silk road can increase social welfare while more successful trades are accomplished and most prosumers in the market are involved.

8.3.1 Modeling Knowledge Trades

More formally, the possible knowledge trades among prosumers can be modeled as a directed graph $G = (V, E)$. We start by restricting each prosumer only supplies one subject and demands another subject of knowledge (single-supply-single-demand, SSSD), but we will also examine the multi-supply-multi-demand (MSMD) mode later. First, for each prosumer, construct one vertex v. Next, from one prosumer v_i to another v_j, add a weighted edge e_{ij} if v_i can supply the knowledge demanded by v_j. The weight w_{ij} is assigned to e_{ij} according to the gain v_j can have from the knowledge supplied by v_i. In our model, it is the TF-IDF weight for computing the relevance between the demanded knowledge described in the epistemology of v_i and the supplied knowledge described in the epistemology of v_j. However, the weight can be a combination of many factors that can measure the quality of knowledge, such as the expertise and reputation of the supplier. A possible exchange is represented as a cycle c in this graph, with each prosumer in the cycle obtaining the knowledge of the next prosumer. The sum of all edge weights of a cycle c is the cycle weight w_c.

A knowledge trading silk road is a collection of cycles, which has the maximum cardinality (i.e., it includes the most vertices, that is to say, the most prosumers can contribute and acquire knowledge). In our model, the cycles are edge-disjoint (no two of

them have an edge in common) and not necessarily vertex-disjoint (no two of them have a vertex in common). This is due to the character of knowledge trading, i.e., one can exchange the same knowledge with others for multiple times while a real item can be exchanged only once. This was illustrated in the following saying that is credited to George Bernard Shaw: "If you have an apple and I have an apple and we exchange apples then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas." Nevertheless, this model can be adapted to suit real-item exchange or user's preference (e.g., prosumers set only receive knowledge from or afford knowledge to one prosumer), if we restrict that the cycles are vertex-disjoint. This restriction does not affect the analytical results of this model.

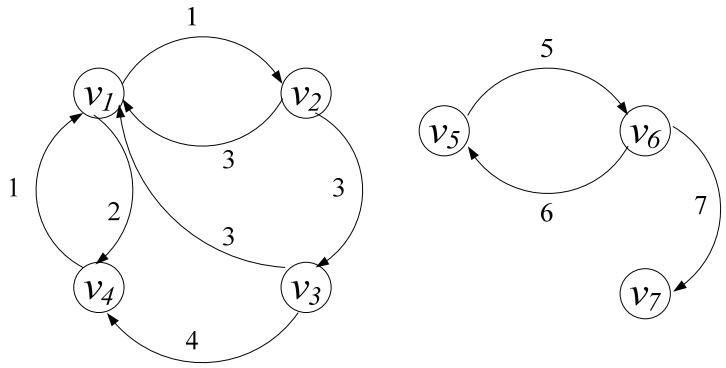

Figure 8.1 Knowledge Trading Example

Figure 8.1 illustrates an example of knowledge exchange with 7 prosumers, $\{v_1, v_2, \ldots v_7\}$, in which all edges are labeled with different weights. There are 5 cycles in the graph, $c_1 = (v_1, v_2)$, $c_2 = (v_1, v_2, v_3)$, $c_3 = (v_1, v_2, v_3, v_4)$, $c_4 = (v_1, v_4)$, and $c_5 = (v_5, v_6)$. Obviously, v_7 is not included in any cycles probably because nobody has demanded the knowledge

supplied by the prosumer. Then we can find two knowledge silk roads, $R_1 = \{c_2, c_4, c_5\}$ and $R_2 = \{c_3, c_5\}$, each is a collection of cycles that cover the most vertices of G.

8.3.2 Social Welfare Maximizing

Let the weight of the knowledge silk road is the sum of its cycle weights. The silk road with maximum weight is the social welfare maximizing silk road, which means if knowledge is traded along this route, the total income (the knowledge that is relevant to what has been required) of the prosumers in the social search society is maximized. For example, in Figure 1, we can get the weights of S_1 and S_2 are (1+3+3)+(1+2)+(5+6)=21 and (1+3+4+1)+(5+6)=20 respectively. Therefore S_1 is the social welfare maximizing silk road. However, if the weight w_{31} is changed from 4 to 6, then S_2 will have the maximum weight. This is because that the income of v_4 will increase more (from 2 to 6) than the income of v_1 will reduce (from 3 to 1), and the social welfare will be maximized in S_2.

In the incentive mechanism for a social search community, the objective is to find the social welfare maximizing silk road which consists cycles in the graph. Further, we must consider additional constraints on the set of cycle. The size or length of the cycles is a natural constraint for the following reasons. First, the knowledge trading in a shorter cycle is more efficient because the trading must be approved by all participants and it will take time if prosumers are not simultaneously online (e.g., due to the time difference). Second, fewer prosumers will be affected if a shorter cycle fails to trade knowledge when any prosumers drop out of the cycle accidentally (e.g., due to changes of requirements). Third, prosumers could be clear about the authenticity of knowledge in shorter cycles, especially in a 2-edge cycle since the two prosumers are both aware that they should trust

each other and exchange authentic knowledge. The constraint of cycle-length has also been applied in existing online real item barter systems. For example, Swap.com[21] (formerly SwapTree), which enables users trade certain stuff such as books, CDs, DVDs, and videogames with other users, can set up trades with up to four people.

Consequently, we focus on studying the problem of finding the maximum-weight knowledge trading silk road consisting of cycles with maximum length at some constant K. This problem actually can be formulated as a variant of the *cycle covering problem* (CCP) in graph theory. Similar to the *traveling salesman problem*, where the goal is to find a Hamiltonian cycle of minimum or maximum weight, this problem can also be proved to be NP-complete, as other variants of CCP [141] (e.g., the constrained version of the *Chinese Postman Problem* [57]). Therefore, approaches based on approximation are generally applied to solve the problem. However, in order to avoid dissatisfaction of trading opportunities caused by the loss of optimality, we aim at an optimal solution based on an *integer linear programming* (ILP) formulation, which is the technique for the optimization of a linear objective function. ILP formulation allows one to add additional constraints to the problem, and to model a number of variations on the objectives. For example, long cycles can be given smaller weights than short cycles if prosumers prefer being involved in a trade with fewer participants.

Now we consider the MSMD mode. Construct vertex s_{ij} for the jth knowledge supply of prosumer i, and vertex d_{ij} for the jth knowledge demand of prosumer i. If we add a zero-

---

[21] http:// www. swap.com

weighted edge from d_{ij} to s_{ik} (dash lines in Figure 8.2), we can see that prosumer i and prosumer j can accomplish a trade if there is a cycle covering their supply and demand vertices (e.g., $c = (d_{11}, s_{12}, d_{21}, s_{21})$ in Figure 8.2). Therefore the MSMD mode is similar to the SSSD mode and can be resolved with the same method.

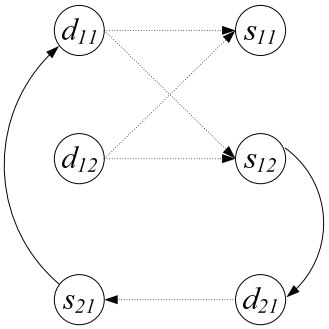

Figure 8.2 The multi-supply-multi-demand mode

8.3.3 Cycle Formulation

In this section, we concentrate on the solution of the maximum-weight silk road problem. A formulation of this problem can be considered as an ILP with one variable for each cycle.

When cycles have length at most 2, this problem can be converted into the classical *assignment problem*, which consists of finding a maximum weight matching in a weighted bipartite graph, and can be solved by a reduction to the perfect matching in the graph. For example, given a knowledge trading graph $G = (V, E)$ (Figure 8.1), a new graph on V (Figure 8.3) can be constructed with a weight w_c edge for each cycle c of length 2. It is easy to see that each matching in the new graph corresponds to a cycle cover by cycles of length 2 in the original graph. Hence, the maximum-weight silk road

problem with K = 2 can be solved in polynomial time by finding a maximum-weight matching.

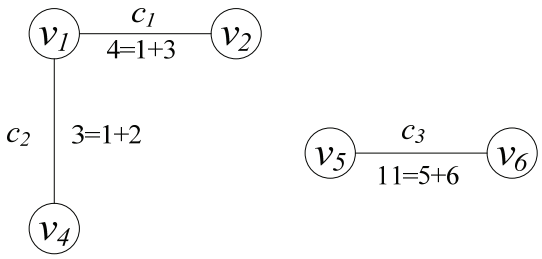

Figure 8.3 Maximum-weight matching with $K = 2$

This formulation can be generalized for arbitrary K. Let $\Gamma_K = \{c_1, c_2, \ldots, c_n\}$ represent the set of all edge-disjoint cycles in G of with length less than or equal to K. Let x_c be a 0-1 variable indicating whether cycle c is selected to be part of the cycle cover or not. Let a_{ce} be 1 if edge e is covered by cycle c and 0 otherwise. Then the maximum-weight cycle cover by Γ_K cycles can be found with the following ILP:

$$\max \sum_{c \in \Gamma_k} w_c x_c$$

$$\text{subject to: } \sum_{c:e \in c} a_{ce} x_c \leq 1 \quad \forall e \in E$$

$$\text{with } x_c \in \{0,1\} \quad \forall c \in \Gamma_k$$

The ILP can be efficiently solved using tree search techniques, including branch-and-bound, branch-and-cut and other branching algorithms. In our approach, the branch-and-bound algorithm is adopted, which is described as follows:

Step 1: Initialize a list P of problems to solve. A problem called the root problem is first included into it. The root problem is the linear programming relaxation, which is a linear program that replaces the constraint in the ILP (i.e. each variable must be 0 or 1) by a weaker constraint (i.e. each variable belong to the interval [0, 1]).

Step 2: If $P=\varnothing$, then the best known feasible solution (silk road) is optimal. Otherwise, choose a problem p and delete it from the list.

Step 3: (i) Solve the linear relaxation of p. If the solution is integral, then update the best known integral solution and the best known solution value eventually and return to *Step 2*.

(ii) If the value of the objective function exceeds that of the best known feasible solution, return to *Step 2*.

(iii) Otherwise, partition the current problem b into two new problems by using some linear inequality, and add them to P. This can be done by choosing a variable with a current fractional value $\bar{x}_{ij}$, then imposing $x_{ij} \geq 1$ in one problem and $x_{ij} \leq 0$ in the other. All feasible solutions of b are contained in the union of the feasible (integral) solutions to each of the two problems. Next, return to *Step 2*.

Figure 8.4 The branch-and-bound algorithm

8.4 Experiments

The goal of the experiments reported on in this section was to verify the feasibility of the knowledge trading silk road as an incentive mechanism for social search, and

demonstrate the effectiveness and efficiency of our algorithm for the solution of the maximum-weight silk road problem.

Our dataset was collected from *Yahoo! Answers*, a social search site in the form of a QA system. We used the *Yahoo! Answers* API[22] to retrieve the questions and answers in different categories from numerous *Yahoo!* users. In order to simulate the experience of a collection of prosumers from a social search community, we only focused on the questions in *Yahoo! New Zealand Answers*. There are, in total, 92,112 questions with at least one answer from April 2008 to March 2011.

We then constructed the knowledge trading graph. However, in QA systems, there is actually no incentive mechanism described in this chapter, i.e., prosumers will give an abridgment or snippet view of the knowledge they can supply in advance (rather than providing an answer to a given question), and the mechanism can find the supply-demand pairs based on their relevance (e.g., TF-IDF score of the descriptions in their epistemologies). Therefore, we first matched the supply-demand simply based on their Category id, that is, if a user asked a question in a category and another user provided an answer in the same category, one edge between them was added with weight 1. This is reasonable since the answerer in a category might be an expert in that field and probably could supply related knowledge demanded by the questioner. Further, if there are common terms in the question and the answer, the edge was assigned with weight 2. We have got 109,027 such supply-demand pairs in our dataset. We also filtered out users only

---

[22] http:// developer.yahoo.com/answers

asked or answered questions because they could not be involved in any knowledge trade. Finally, we got 1,430 users and 23,210 questions.

Table 8.1 Characteristics of the knowledge trading graph

Length of cycles	Number of cycles	Average weights of maximum-weight cycle covers	Coverage
K=2	2275	324.2	13.1%
K=3	2618	360.9	14.4%
K=4	2939	383.6	15.3%
K=5	3287	402.7	15.9%
K=6	3506	413.3	16.2%
K=7	3718	420.7	16.4%
K=8	3841	426.5	16.5%

Table 8.1 shows characteristics of the knowledge trading graph for our dataset. We can see that there were more than 2200 cycles in the graph if the length of cycles was limited to no more than 2, and more than 3800 cycles could be found with the constraint K=8. This has confirmed the feasibility of knowledge trading and the incentive mechanism in a social search community, because a prosumer can obtain the required knowledge through trading the possessed knowledge with other prosumers.

However, the coverage of the maximum-weight cycle covers was low in the dataset, which means that there would be a large amount of knowledge supplies and demands not included in the covers. For example, only 13.1% knowledge supplies and demands could be involved in the maximum-weight cycle covers, if no trades except two-way exchanges were allowed (K=2). This is not strange due to the lack of our incentive mechanism in the

QA system, where users would not and could not trade the knowledge they possessed for the knowledge they required. The knowledge supplies and demands were unbalanced in the market as most users asked much more questions than answers they provided. Therefore fewer vertices in the graph could be included in the maximum-weight cycle covers. On the contrary, in our *Baijia* system, prosumers are aware that the more knowledge they contribute, the greater the opportunity of obtaining the knowledge they required. Consequently there would be more cycles in the graph and the coverage of maximum-weight cycle covers would be improved.

Further, Table 8.1 also shows that the average weights of maximum-weight covers (the maximized social welfare) by cycles with length no more than 5 are almost as large as the average weights of covers by cycles with K=8. Therefore it is possible to reduce the computational complexity by restricting the length of cycles to 5, while still getting close to the optimal social welfare.

To analyze the performance of our branch-and-bound algorithm for the solution of the maximum-weight silk road problem, we have compared it with two other algorithms. The first algorithm is a lowest-first *branch-and-cut* procedure [7], where the LP relaxation is initialized by taking all the constraints present in the last LP solved at the parent node at each node of the branching tree. The second is a *local search* algorithm [122] that wanders through the space of feasible solutions, which moves from one feasible solution to another one nearby measured by Hamming distance at each step. The experiments were carried out on a Dell Precision Workstation T3400, with Intel Core2 Quad CPU

Q9450 @ 2.66GHz and 8 GB of RAM, running Windows XP Professional (version 2002).

Table 8.2 Computation times of the three algorithms

Length of cycles	Branch-and-bound	Branch-and-cut	Local search
K=2	9.2	15.2	53.7
K=3	14.5	25.4	92.0
K=4	23.8	35.1	149.2
K=5	36.4	50.6	197.1
K=6	49.7	79.3	240.8
K=7	70.1	110.8	312.3
K=8	93.9	158.4	398.7

In Table 8.2, we report the computation times (in CPU seconds on the Dell T3400) for various cycle length constraints of the three algorithms we described above. On all cycle length constraints, Table 8.2 show that the performance of our branch-and-bound algorithm is better than other two algorithms. However, it is worth exploring advanced heuristic techniques to further improve the efficiency of the solving the maximum-weight silk road problem, and make the proposed incentive mechanism more feasible for online social search systems.

8.5 Summary

Social search is gaining increasing momentum as EIS tasks are constantly on the rise and more people are joining the online social networking force. However, analysis on the

dataset we collected from a well-known social search system - *Yahoo! Answers* - has discovered a significant imbalance between the number of people who seek help and the number of volunteers who offer help. Most people have asked a lot of questions but offered few or no answers to the questions asked by others and on average each participant asks a lot more questions than the number of answers they have actually provided to others. This discovery suggests that the points-based incentive mechanism built on the free trading model adopted by most QA systems is not good enough to encourage people to contribute their answers.

In this chapter we present the non-monetary incentive mechanism for social search systems, including the online silk road for the epistemology trading and an efficient algorithm to achieve the maximum social welfare based on an *integer linear programming* formulation, where the total of weights for the supply-demand pairs in the knowledge market can be maximized. This mechanism encourages users to contribute high-quality content and trade it with others, thus can make a social search website attractive and sustainable. We apply the proposed approach on the *Yahoo! Answers* data to simulate the epistemology trading in a social search community. The results of experiments show that the incentive mechanism is feasible and our algorithm is more efficient than other algorithms.

An imminent future work is to collect feedbacks from users based on our prototype system. We can then conduct usability tests to measure the balance (or imbalance) between the supplies and demands, and testify whether this mechanism encourages more and better contributions compared to the mechanisms adopted by existing QA systems.

We also plan to incorporate the copyright model (as used in YouTube) into our incentive mechanism for epistemology contribution. Further, we will explore centralized portals for ranking all epistemology providers to ensure the quality of epistemologies.

Chapter 9

Conclusions and Future Work

In this chapter, we conclude the thesis by summarizing our contributions, and propose some possible directions for future research work.

9.1 Conclusions

This thesis aims to provide a comprehensive solution to make use of the collective wisdom effectively by leveraging IR, knowledge discovery, and social network analysis techniques for EIS. To this purpose, we propose an EPISOSE framework and develop several approaches to tackle various challenging issues in the components of the framework including epistemology generation, epistemology search, epistemology editing and refining, and epistemology services. The major achievements and contributions are concluded in the following.

First of all, a novel epistemology-based social search framework is proposed for EIS, where search epistemologies – aggregated and well-structured information packages derived from successful search processes contributed by a mass of prosumers – are effectively shared, reused, and refined by others with same or relevant search interests or goals. This framework can be applied to the design and implementation of a range of

social search systems with different strategies and algorithms, and we have designed and implemented a prototype system *Baijia* with the guidance of the framework.

It is important to automatically generate epistemologies from prosumers' interactions with the system. To tackle this issue, a novel probabilistic topic model with social tags has been proposed to discover the latent semantic relationships between query terms and Web pages. The epistemology correlated with an EIS task is generated through modeling the prosumer's search process. The generated epistemology contains queries, URLs and tags that are relevant to the search goal. Moreover, related queries are discovered based on the similarity of probability distributions over topics, and the retrieved URLs can be ranked based on the similarity of probability distributions of the URLs and the queries.

Epistemology retrieval is a critical issue in the social search solution, because prosumers usually have difficulties in formulating proper keywords in EIS processes. In contrast to previous techniques revolving around mining existing queries that are most similar to a given query or providing most dissimilar results for a given query, we propose a novel social-interest-directed technique which can not only suggest highly diverse queries that are yet closely related to a given query, but also provide diverse epistemologies for prosumers with vague information needs. Social interest is discovered by employing the KPCA on the related queries and epistemologies contributed by enormous prosumers. An algorithm is developed to incorporate the social interest with a random walk on the query-URL bipartite graph.

The epistemologies in our solution can be further edited and refined to help prosumers achieve their final search goal. We propose the IPOD approach to help consumers acquire

the non-existent information though a consumer-led interactive search process, where invited information providers from relevant social networks jointly edit and refine the epistemology on the fly to meet the consumer's needs. Prosumers with the same or similar search interests are clustered in the social networks built based on the epistemologies. The system thereby can identify potential providers in the social network of a consumer, and use a pre-structured epistemology to represent a consumer's information needs so that the prosumers can edit and refine the epistemology in an interactive search process.

Social networking is important in supporting epistemology sharing between prosumers in the social search community, and trust management in social networks is a paramount issue in the epistemology services component as the epistemology quality is closely related to the trustworthiness of the user who contributes or refines the epistemology. Inspired by the physical and mathematical properties and the power flow study in electrical grids, we propose the credit flow approach to modeling and inferring trust relations in social networks. This approach can construct a web of credit from the interactions in a social network, such as the contribution and evaluation of epistemologies, and infer trust values by making use of both generally agreed reliability and subjective individuality in the network.

An incentive mechanism to encourage users to contribute knowledge is imperative to make a social search website attractive and sustainable. We propose a non-monetary incentive mechanism for social search systems based on the epistemology trading. This approach nurtures social search through the online silk road for modeling the

epistemology trading. Prosumers can trade their possessed knowledge for required knowledge without using the virtual currency as a medium of exchange. Moreover, we develop an efficient algorithm to achieve the maximum social welfare based on an *integer linear programming* formulation, where the total of weights for the supply-demand pairs in the knowledge market can be maximized.

9.2 Future work

Today a brilliant new era of social search is upon us. In recent years, latest research findings and paradigms around this topic, including social bookmarking, personalized ranking, question answering, and collaborative searching are emerged continuously. One of the ongoing efforts most immediately connected to this thesis is the attempt to incorporate social network into the search results of commercial SEs such as *Google* [32] and *Bing* [107]. However, unlike the framework proposed in this thesis, all these SEs are limited to the information that the user allows them to access. A full comparison of the work in this thesis with existing frameworks and systems is shown in Appendix B.

Although this thesis has presented a comprehensive and promising solution of the epistemology-based social search for EIS, there are still some open issues that need to further explore in future work.

9.2.1 New Evaluation Paradigms for Exploratory Information Seeking Systems

The deployment of EIS systems opens up new opportunities for their evaluation. In this thesis, we have discussed the evaluation of such systems by using some widely used metrics for IR or some specific metrics for EIS. However, for our social search solution supporting EIS, the manipulation of epistemologies by thousands, if not millions, of users, makes it difficult to evaluate the effectiveness of EIS in a laboratory setting. The system should also be developed with the ability to monitor the robustness of the system if it is subject to malicious rank manipulation. We can design new evaluation paradigms for EIS systems as we believe that more effective evaluation methodologies will improve the quality of these systems that reach users. For example, we can evaluate the system with simulations using the data extracted from TREC interactive search logs, and we can also evaluate it in the context of an interactive collaborative information seeking activity involving real users.

9.2.2 Domain Specific Social Search in Virtual Communities

As opposed to general-purpose search, vertical or domain-specific search can support specific unique user tasks and achieve greater precision by leveraging domain knowledge and reducing search scope [15]. A future direction is to apply the epistemology-based social search to a specified application domain.

For example, as it is increasingly important for people to be more proactive in maintaining their health but current health information SEs are not good enough for

average citizens to explore health information on the Web, we can apply the technology developed in this thesis to a specialized exploratory health information search system. We can use query logs from existing specialized health information SEs such as WebMD and PubMed, and health-related query records from existing general-purpose SEs to train the topic model described in Chapter 4, or use health-related criteria to find the best collaborators for collaborative epistemology editing described in Chapter 6. Further, the key epistemology services developed in Chapter 7 and Chapter 8 can also be built upon virtual communities for creating and sharing health information.

9.2.3 Revenue Models for Social Epistemology Economy

Although Chapter 8 of this thesis has described an incentive mechanism to encourage users in participating social search through epistemology trading, it is still an open question whether a social search system such as *Baijia* could be sustainable with existing revenue models, including onsite advertising and subscription fees. In fact the epistemology is a kind of product of social labor and can be the thrust of a new form of economy – the social epistemology economy. It is produced and circulated in the frame of economic constraints. Therefore new revenue models for epistemology-based social search websites might come from this aspect. For example, if a product is recommended in an epistemology of seeking for related information, both the prosumers and the website can get revenue from the manufacturer.

The challenge here is that to differentiate real users from astroturfing can be very complex, probably involving the privacy issues. How to develop robust and efficient

mechanisms for preserving the “copyrights” (valuable experience or knowledge) of real users is a future research direction.

Bibliography

[1] Agichtein, E., Brill, E., and Dumais, S. 2006. Improving web search ranking by incorporating user behavior information. In SIGIR ’06, ACM Press, pp. 19–26.

[2] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. 2009. Diversifying search results. Proc. The Second ACM International Conference on Web Search and Data Mining (WSDM 09), ACM Press, pp. 5-14.

[3] Alodhaibi, K., Brodsky, A., and Mihaila, G. A. 2011 A Randomized Algorithm for Maximizing the Diversity of Recommendations. Proc. of the 44th Hawaii International Conference on System Sciences (HICSS’11). IEEE Computer Society, pp. 1-10.

[4] Anderson, C. 2006. The long tail: Why the future of business is selling less of more. New York: Hyperion.

[5] Anheier, H. 2003. Movement Development and Organizational Networks: The Role of ‘Single Members’ in the German Nazi Party, 1925-30. Social Movements and Networks: Relational Approaches to Collective Action. New York, Oxford University Press, pp.49-76.

[6] Armin, H. 2004. Query Expansion Methods for Collaborative Information Retrieval. Technical University of Kaiserslautern, Germany, Ph.D. thesis.

[7] Ascheuer, N., Jünger, M., Reinelt, G. 2000. A branch & cut algorithm for the asymmetric Traveling Salesman Problem with precedence constraints. Computational Optimization and Applications 17(1), pp. 61–84.

[8] Aslam, J. and Montague, M. 2001. Models for metasearch. In SIGIR ’01, ACM Press, pp. 276–284.

[9] Baeza-Yates, R., Hurtado, C. and Mendoza, M. 2004. Query clustering for boosting web page ranking. In AWIC, volume 3034 of Lecture Notes in Computer Science, pp. 164–175.

[10] Baeza-Yates, R., Hurtado, C., and Mendoza, M. 2004. Query recommendation using query logs in search engines. In: Proc. of International Workshop on Clustering Information over the Web (ClustWeb, in conjunction with EDBT), pp. 588-596.

[11] Baeza-Yates, R. and Ribeiro-Neto, B. 1999. Modern Information Retrieval. ACM Press, ISBN: 020139829.

[12] Baeza-Yates, R., Tiberi, A. 2007. Extracting semantic relations from query logs. In KDD'07, Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM Press, pp. 76-85.

[13] Bao, S., Wu, X., Fei, B., Xue, G., Su, Z., Yu, Y. 2007. Optimizing Web Search Using Social Annotation. In WWW'07, Proceedings of the 16th World Wide Web Conference, ACM Press, pp. 501-510.

[14] Baraglia, R., Nardini, F. M., Castillo, C., Perego, R., Donato, D., and Silvestri, F. 2010. The effects of time on query flow graph-based models for query suggestion. In: Proc. of Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO 10), pp. 182-189.

[15] Battelle, J. 2005. The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed Our Culture. Portfolio.

[16] Beeferman, D., and Berger, A. 2000. Agglomerative clustering of a search engine query log. Proc. The 6th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD 00), ACM press, pp. 407–416.

[17] Belkin, N. J. and Croft, W. B. 1992. Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 35(12), pp. 29–38.

[18] Berger, A. and Lafferty, J. Information retrieval as statistical translation. 1999. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, pp. 222−229.

[19] Berners-Lee T, Hendler J. April, 2001. Scientific publishing on the Semantic Web. Nature [serial online]. 26, 410(6832), pp. 1023–1024.

[20] Bhogal, J., Macfarlane, A., and Smith, P. 2007. A review of ontology based query expansion. Information Processing & Management. 43(4): 866–886.

[21] Bian, J., Liu, Y., Agichtein, E., and Zha, H. 2008. Finding the right facts in the crowd: factoid question answering over social media. In WWW'08, Proceedings of the 17th World Wide Web Conference, ACM Press, pp. 467-476.

[22] Billerbeck, B., Scholer, F., Williams, H. E., and Zobel, J. 2003. Query expansion using associated queries. In CIKM'03, Proceedings of the 12th ACM Conference on Information and Knowledge Management, ACM Press, pp. 2-9.

[23] Blanzieri, E., Giorgini, P., Massa, P., Recla, S. 2001. Implicit culture for multi-agent interaction support. In CoopIS'2001, Proceedings of the 9th International Conference on Cooperative Information Systems, Springer, pp. 27–39.

[24] Blei, D. M. and Lafferty, J. 2006. Correlated topic models. In NIPS'06: Proceedings of Advances in Neural Information Processing Systems.

[25] Blei, D. M. and McAuliffe, J. D. Supervised topic models. 2007. In NIPS'07: Proceedings of Advances in Neural Information Processing Systems.

[26] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. Journal of Machine Learning Research, pp. 993–1022.

[27] Bookstein, A. 1983. Information retrieval: A sequential learning process. Journal of the American Society for Information Sciences (ASIS), 34(5), pp. 331-342.

[28] Broder, A. 2002. A taxonomy of web search. SIGIR Forum, 36(2).

[29] Buckley, C., and Voorhees, E. M. 2000. Evaluating evaluation measure stability. In SIGIR'00, ACM Press, pp. 33−40.

[30] Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., and Li, H. 2008. Context-aware query suggestion by mining click-through and session data. In: Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 08), ACM press, pp. 875-883.

[31] Carbonell, J., and Goldstein, J. 1998. "The use of MMR, diversity-based reranking for reordering documents and producing summaries." Proc. The 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 98), ACM Press, pp. 335-336.

[32] Cassidy. M. An update to Google social search. 2011. The official Google Blog, Feb. 17, 2011.

[33] Castronova, E. October 2006. A Cost-Benefit Analysis of Real-Money Trade in the Products of Synthetic Economies. Info, Vol.8, No.6.

[34] Chris Sherman and Gary Price. 2001. The invisible web: Uncovering information sources search engines can't see.

[35] Chen, H. and Karger, D.R. 2006. Less is more: probabilistic models for retrieving fewer relevant documents. Proc. The 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 06), ACM Press, pp. 429-436.

[36] Chi, Ed H. 2009. Information Seeking Can Be Social, *Computer*, 42, 3, pp. 42-46.

[37] Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I. 2008. Novelty and diversity in information retrieval evaluation. Proc. The 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 08), ACM Press, pp. 659-666.

[38] Craswell, N. and Szummer, M. 2007. Random walks on the click graph. Proc. The 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 07), ACM Press, pp. 239-246.

[39] Cui, H., Wen, J. R., Nie, J. Y., and Ma, W. Y. 2003. Query expansion by mining user logs. IEEE Transaction of Knowledge Data Engineering, 15(4), pp. 829–839.

[40] Davenport, H. T. & Prusak, L. 1998. Working knowledge: How organizations manage what they know. Boston, MA: Harvard Business School Press.

[41] 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), pp. 391–407.

[42] Dou, Z., Song, R., Wen, J. 2007. A large-scale evaluation and analysis of personalized search strategies. In Proc. WWW'07, ACM Press, pp. 581-590.

[43] Douceur, John R. 2002. The Sybil Attack. *International workshop on Peer-To-Peer Systems*.

[44] Evans, B. M., Chi, E. H. 2008. Towards a model of understanding social search. In CSCW'08, Proceedings of the ACM Conference on Computer-Supported Cooperative Work, ACM Press, pp. 485-494.

[45] Fonseca, B. M., Golgher, P., Possas, B., Ribeiro-Neto, B., and Ziviani, N. 2005. Concept-based interactive query expansion. In: Proc. of the 14th Conference on Information and Knowledge Management (CIKM 05), ACM Press, pp. 696-703.

[46] Fu, W. 2008. The microstructures of social tagging: a rational model. In Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, San Diego, *USA*, ACM Press, pp. 229-238.

[47] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. 1987. The Vocabulary Problem in Human-System Communication. Communications of the ACM, 30(11), pp. 964–971.

[48] Fuxman, A., Tsaparas, P., Achan, K. and Agrawal, R. 2008. Using the wisdom of the crowds for keyword generation. Proc. The 17th international conference on World Wide Web (WWW 08), ACM Press, pp. 61-67.

[49] Gianoutsos, S. and Grundy, J. 1996. Collaborative work with the World Wide Web: adding CSCW support to a web browser. In Proceedings of OZ-CSCW, pp. 14–21.

[50] Glance, N. Community search assistant. 2001. In: Proc. of the 6th International Conference on Intelligent User Interfaces (IUI 01), ACM Press, pp. 91-96.

[51] Goffman, W. 1964. On relevance as a measure. Information Storage and Retrieval, Vol. 2, pp. 201-203.

[52] Golbeck J. 2005. Computing and Applying Trust in Web-based Social Networks. PhD thesis, University of Maryland.

[53] Gollapudi, S., and Sharma, A. 2009. An axiomatic approach for result diversification. Proc. The 18th International Conference on World Wide Web (WWW 09), ACM Press, pp. 381–390.

[54] Gotta, M. 2008. Analysis of Social Networks: Telling Old Stories in New Ways. 3th April 2008. Available from: http://mikeg.typepad.com/perceptions/2008/04/analysis-of-soc.html.

[55] Greenberg, S. and Roseman, M. 1996. Groupweb: a www browser as real time groupware. In Proceeding of the14th annual SIGCHI Conference on Human Factors in Computing Systems (CHI'96), pp. 271–272.

[56] Griffiths, T. L. and Steyvers, M. 2004. Finding scientific topics. In Proceeding of the National Academy of Sciences, pp. 5228–5235.

[57] Guan, M. 1962. Graphic programming using odd and even points. Chinese Mathematics, 1, pp. 273–277.

[58] Guha R. 2003. Open rating systems, Technical report, Stanford University, CA, USA.

[59] Guha R., Kumar R., Raghavan P., and Tomkins A. 2004. Propagation of trust and distrust. In Proc. of WWW'04, pp. 403–412.

[60] Guha, R. V., McCool, R., and Miller, E. 2003. Semantic search. In Proceedings of the 12th International WorldWide Web Conference (WWW'03), pp. 700–709.

[61] Guo, J., Chow, A., Gong, Z., Sun, C. 2009. Virtual Wealth Realization in Virtual and Real Worlds, IEEE International Conference on e-Business Engineering, ICEBE '09, pp. 85-94.

[62] Hansen, P. & Järvelin, K. 2005. Collaborative information retrieval in an information-intensive domain. Information Processing & Management 41(5), pp. 1101-1119.

[63] Hartigan, J. and Wong, M. 1979. A k-means clustering algorithm. Applied Statistics. 28(1979), pp. 100-108.

[64] Haveliwala, T., Kamvar, S., and Jeh, G. June, 2003. An analytical comparison of approaches to personalizing PageRank. Technical Report, Stanford University.

[65] Herlocker, J.L., Konstan, J.A., Borchers, A. and Riedl, J. 1999. An Algorithmic Framework for Performing Collaborative Filtering. In SIGIR '99, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, pp. 230-237.

[66] Heymann, P., Koutrika, G. and Garcia-Molina, H. 2008. Can social bookmarking improve web search? In Proceedings of the 1st ACM International Conference on Web Search and Data Mining, Stanford, USA, ACM Press, pp. 195-205.

[67] Hiemstra, D. 2001. Using language models for information retrieval. Ph.D. Thesis, Centre for Telematics and Information Technology.

[68] Hofmann, T. 1999. Probabilistic latent semantic indexing. In Proceedings of SIGIR'99, ACM Press, pp. 50–57.

[69] Horowitz, D. and Kamvar, S. D. 2010. The anatomy of a large-scale social search engine. In Proc. WWW '10. ACM Press, pp. 431-440.

[70] Hörster, E., Lienhart, R., and Slaney, M. 2008. Image retrieval on large-scale image databases. In CIVR'08, pp. 17–24.

[71] Huggins, R., Johnston, A. & Steffenson, R. 2008. Universities, knowledge networks and regional policy Cambridge Journal of Regions, Economy and Society, 1, 2, pp. 321-340.

[72] Irwin, D., Chase, J. Grit, L. and Yumerefendi, A. 2005. Self-recharging virtual currency, Proceedings of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems, Philadelphia, Pennsylvania, USA: ACM Press, pp. 93-98.

[73] Jansen, B. J., Campbell, G., and Gregg, M. 2010. Real time search user behavior. Ext. Abstracts CHI 2010. ACM Press, pp. 3961-3966.

[74] Jarvelin, K. and Kekalainen, J. 2000. IR evaluation methods for retrieving highly relevant documents. In SIGIR'00, ACM Press, pp. 41–48.

[75] Jaschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. 2007. Tag Recommendations in Folksonomies. Lecture Notes in Artificial Intelligence, 4702, pp. 506-514.

[76] Jeh, G. and Widom, J. 2003. Scaling personalized web search. In WWW'03, Proceedings of the 12th International World Wide Web Conference, ACM press, pp. 271-279.

[77] Joachims, T. 2002. Optimizing search engines using clickthrough data. In KDD'02, ACM Press, pp. 133–142.

[78] Jones, R., Rey, B., Madani, O., and Greiner, W. 2006. Generating query substitutions. In: Proc. of the 15th International Conference on World Wide Web (WWW 06), ACM Press, pp. 387-396.

[79] Jøsang, A. 1999. An Algebra for Assessing Trust in Certification Chains. In *Proc. of the* Network and Distributed Systems Security (NDSS'99) Symposium, The Internet Society.

[80] Kamerer, Y., Nairn, R., Pirolli, P., and Chi, E. H. 2009. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proc. CHI'09*, ACM Press, pp. 625-634.

[81] Kamvar, S. D., Schlosser, M. T., and Garcia-Molina, H. 2003. The eigentrust algorithm for reputation management in p2p networks. In *Proc. of WWW'03*, ACM Press, pp. 640–651.

[82] Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM 46, 5(1999), pp. 604-632.

[83] Knoke, D. and Yang, S. 2008. Social Network Analysis. 2nd Edition ed. Quantitative Applications in the Social Sciences. SAGE Publications, pp. 144.

[84] Kollock, P. and M. Smith. 1996. Managing the Virtual Comments: Cooperation and Conflict in Computer Communities. Computer-Mediated Communication, edited by Susan Herring. Amsterdam, John Benjamins.

[85] Koren, J., Zhang, Y., Liu, X. 2008. Personalized interactive faceted search. In WWW'08, Proceedings of the 17th World Wide Web Conference, ACM Press, pp. 477-485.

[86] Kraft, R. and Zien, J. 2004. Mining anchor text for query refinement. In: Proc. of the 13th International Conference on World Wide Web (WWW 04), ACM Press, pp. 666-674.

[87] Kubica, J., Moore, A., Schneider, J., and Yang, Y. 2002. Stochastic Link and Group Detection. In Proceedings of AAAI Workshop on Link Analysis.

[88] Kwon, K., Cho, J. and Park, Y. 2009. Multidimensional credibility model for neighbor selection in collaborative recommendation. Expert Systems with Applications, 36 (3), pp. 7114-7122.

[89] Lafferty, J., Zhai, C. 2001. Document language models, query models, and risk minimization for information retrieval. In Research and Development in Information Retrieval, pp. 111–119.

[90] Lavrenko, V. and Croft, W. B. 2001. Relevance based language models. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, pp. 120−127.

[91] Lee, Y-J. May 2005. VisSearch: A Collaborative Web Searching Environment. Computers & Education, 44(4), pp. 423-439.

[92] Leuken, R. van, Garcia, L., Olivares, X., Zwol, R. van. 2009. Visual Diversification of Image Search Results. Proc. The 18th international conference on World Wide Web (WWW 09), ACM Press, pp. 341-350.

[93] Levien, R. 2003. Advogato Trust Metric. PhD thesis, UC Berkeley, USA.

[94] Li, X., Wang, Y. Y., and Acero, A. 2008. Learning query intent from regularized click graphs. Proc. The 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'08), ACM Press, pp. 339–346.

[95] Linden, G., Smith, B., and York, J. 2003. Amazon.com recommendations. IEEE Internet Computing 7, no. 1, pp. 76-80.

[96] Liu, S., Liu, F., Yu, C., and Meng, W. 2004. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In: Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 04), ACM Press, pp. 266-272.

[97] Liu, X., Datta, A., Rzadca, K., and Lim E. P. 2009. Stereotrust: a group based personalized trust model. In *Proc. of CIKM'09*, pp. 7–16

[98] Lux, M., Dosinger, G. 2007. From folksonomies to ontologies: employing wisdom of the crowds to serve learning purposes. International Journal of Knowledge and Learning 3(4/5), pp. 515–528.

[99] MacKay, David. 2003. An Example Inference Task: Clustering. Information Theory, Inference and Learning Algorithms. Cambridge University Press, pp. 284–292.

[100] Maheswaran, M., Cheong, T. H., and Ghunaim, A. 2007. Towards a Gravity-Based Trust Model for Social Networking Systems. In Proc. IEEE ICDCS Workshop on Trust and Reputation Management.

[101] Manchala, D.W. 1998. Trust metrics, models and protocols for electronic commerce transactions. In Proc. of ICDCS'98, pp. 312 - 321.

[102] Manning, C., Raghavan, P., and Schütze, H. 2008. Introduction to Information Retrieval. Cambridge University Press.

[103] Marchionini, G. April 2006. Exploratory search: From finding to understanding. Communications of the ACM, vol. 49, pp. 41-46.

[104] Martinez, M. 2003. High Attrition Rates in E-Learning: Challenges, Predictors, and Solutions. The E-Learning Developer's Journal, July 14, 2003, 9 pages.

[105] Massa, P. and Avesani, P. 2005. Controversial users demand local trust metrics: An experimental study on epinions.com community. In *AAAI 2005*, pp. 121–126.

[106] McAdam, D. 1988. Freedom Summer. New York, Oxford University Press.

[107] Mehdi. Y. 2011. Facebook friends now fueling faster decisions on Bing. Microsoft Bing Search Blog, May 16, 2011.

[108] Mei, Q., Zhou, D., and Church, K. 2008. Query suggestion using hitting time. In Proc. of the 17th Conference on Information and Knowledge Management (CIKM 08), ACM Press, pp. 469–478.

[109] Menczer, F., Pant, G., and Srinivasan, P. 2004. Topical web crawlers: Evaluating adaptive algorithms. ACM Trans. Internet Technol. 4(4), pp. 378-419.

[110] Merx, M. Nijbof, W. J. 2005. Factors Influencing Knowledge Creation and Innovation in an Organization. Journal of European Industrial Training, 29, 2, pp. 135-147.

[111] Money trouble in second life. 2007. Technology Review, August 2007.

[112] Monique V. Vieira, Bruno M. Fonseca, Rodrigo Damazio, Paulo Braz Golgher, Davi de Castro Reis, Berthier Ribeiro-Neto. 2007. Efficient search ranking in

social networks. In CIKM'07, Proceedings of the 16th ACM conference on Conference on information and knowledge management, ACM Press, pp. 563-572.

[113] Montaner, M., Lopez, B, & Lluis De La, J. 2003. A taxonomy of recommender agents on the Internet. *Artificial Intelligence Review*, 19, pp. 285-330.

[114] Morris, M.R. 2007. Collaborating Alone and Together: Investigating Persistent and Multi-User Web Search Activities, Microsoft Research Technical Report #MSR-TR-2007-11. January 2007.

[115] Morris, M. R. and Horvitz, E. 2007. SearchTogether: an interface for collaborative web search. In *Proc. UIST'07*, ACM Press, pp. 3-12.

[116] Mui, L., Mohtashemi, M., Halberstadt, A. 2002. A computational model of trust and reputation, In Proc. of HICSS 35, IEEE Computer Society, pp. 188–196.

[117] MySpace is the number one website in the U.S. according to hitwise. HitWise Press Release, July, 11, 2006. http://www.hitwise.com/press-center/hitwiseHS2004/social-networking-june-2006.php.

[118] Nigel Shadbolt, Tim Berners-Lee and Wendy Hall. May/June 2006. The Semantic Web Revisited. IEEE Intelligent Systems 21(3) pp. 96-101.

[119] Ogilvie, P., Callan, J. 2001. The effectiveness of query expansion for distributed information retrieval. In CIKM'01, Proceedings of the 10th ACM Conference on Information and Knowledge Management, ACM Press, pp. 183-190.

[120] Özmutlu, H. C. and Çavdur, F. 2005. Application of automatic topic identification on excite web search engine data logs. Information Processing and Management, 41(2005), pp. 1243–1262.

[121] Page, L., Brin, S., Motwani, R., and Winograd, T. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

[122] Papadimitriou, C. H. and Steiglitz, K. 1977. On the complexity of local search for the traveling salesman problem. SIAM Journal on Computing, 6 (1977), pp. 76-83.

[123] Pass, G., Chowdhury, A., and Torgeson, C. 2006. A picture of search. Proc. The 1st International Conference on Scalable Information Systems (Infoscale 06).

[124] Phadke, A. G. and Thorp, J. S. 1998. Computer relaying for power systems. John Wiley & Sons, Inc., New York, NY, USA.

[125] Phelan, O., McCarthy, K., and Smyth, B. 2009. Using twitter to recommend real-time topical news. In *Proc. RecSys '09.* ACM Press, pp. 385-388.

[126] Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., and Back, M. 2008. Algorithmic Mediation for Collaborative Exploratory Search. In SIGIR'08, Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 315-322.

[127] Ponte, J.M. and Croft, W.B. 1998. A language modeling approach to information retrieval. In Proceedings of the 31st Annual International Conference of ACM's Special Interest Group on Information Retrieval, Melbourne, Australia, ACM Press, pp. 275-281.

[128] Porter, M. F. 1997. An algorithm for suffix stripping. Readings in information retrieval, pp. 313–316.

[129] Pujol, J.M., Sangüesa, R., and Bermúdez, J. 2003. Porqpine: A Distributed and Collaborative Search Engine. In WWW'03, Proceedings of the 12th World Wide Web Conference (poster).

[130] Radlinski, F. and Joachims, T. 2005. Query Chains: Learning to Rank from Implicit Feedback, In KDD'05, Proceedings of the 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM Press, pp. 239-248.

[131] Rafiei, D., Bharat, K., Shukla, A. 2010. Diversifying Web search results. Proc. The 19th international conference on World Wide Web (WWW 10), ACM Press, pp. 781–790.

[132] Ravi, S. 2005. Optimal search engine marketing strategy. International Journal of Electronic Commerce 10, 1, pp. 9-25.

[133] Richardson, M., Agrawa, R., Domingos P. 2003. Trust management for the semantic web. In Proc of ISWC'03. Springer Verlag, pp. 351-368.

[134] Richardson, M., and Domingos, P. 2002. Mining knowledge-sharing sites for viral marketing. In *KDD'02*. ACM Press, pp. 61-70.

[135] Rijsbergen, V. 1986. A new theoretical framework for information retrieval. SIGIR'86, ACM Press, pp.194–200,

[136] Rocchio, J. J. 1971. Relevance feedback in information retrieval. In Salton G. (Ed), The SMART Retrieval System. Englewood Cliffs, N.J: Prentice Hall, Inc. pp. 313−323.

[137] Romano, N., Nunamaker, J., Roussinov, D., and Chen, H. 1999. Collaborative Information Retrieval Environment: Integration of Information Retrieval with Group Support Systems. In HICSS'99, Proceedings of the 32nd Annual Hawaii International Conference on System Sciences, IEEE Computer Society, pp. 1-10.

[138] Rose, D. E. and Levinson, D. 2004. Understanding user goals in web search. In WWW'04, ACM Press, pp. 13–19.

[139] Saadat, H. 2002. Power System Analysis, 2nd Edition, McGraw-Hill, New York.

[140] Sahami, M. and Heilman, T. 2006. A Web-based kernel function for measuring the similarity of short text snippets. In: Proc. of the 15th International Conference on World Wide Web (WWW 06), ACM Press, pp. 377-386.

[141] Sahni, S. and Gonzalez, T. 1976. P-Complete Approximation Problems. J. ACM 23, 3 (July 1976): pp. 555-565.

[142] Salton, G. editor. 1971. The SMART Retrieval System—Experiments in Automatic Document Retrieval. Prentice Hall Inc., Englewood Cliffs, NJ.

[143] Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), pp. 513-523.

[144] Sankaralingam, K., Sethumadhavan, S., and Browne, J. 2003. Distributed pagerank for P2P systems. In Proc. of the 12th IEEE International Symposium on High Performance Distributed Computing, IEEE Computer Society.

[145] Santos, R. L. T., Peng, J., Macdonald, C., and Ounis, I. 2010. Exploiting Query Reformulations for Web Search Result Diversification. Proc. The 19th international conference on World Wide Web (WWW 10), ACM Press, pp. 881-890.

[146] Schein A. I., Popescul A., Ungar L. H. 2002. Methods and Metrics for Cold-Start Recommendations, In SIGIR'02, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, pp. 253-260.

[147] Schneiderman, B., Byrd, D. and Croft, W. B. 1997. Clarifying search: A user-interface framework for text searches. D-Lib Magazine, January, 1997.

[148] Scholkopf, B., Smola, A., and Müller, K. R. 1999. Kernel principal component analysis. Advances in Kernel Methods - Support Vector Learning, MIT Press, Cambridge, MA, pp. 327–352.

[149] Scholkopf, B., Smola, A., and Müller, K. July 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, vol. 10, pp.1299–1319.

[150] Shafer, G. 1976. A Mathematical Theory of Evidence. Princeton University Press, Princeton, NJ.

[151] Shankland, S. 2009. Relevance meets the real-time web, http://googleblog.blogspot.com/2009/12/relevance-meets-real-time-web.html.

[152] Sharad Goel, Roby Muhamad, Duncan J. 2009. Watts: Social search in "Small-World" experiments. In WWW'09, Proceedings of the 18th World Wide Web Conference, ACM Press, pp. 701-710.

[153] Shin, D. H. 2008. Understanding purchasing behaviors in a virtual economy: Consumer behavior involving virtual currency in Web 2.0 communities. Interacting with Computers, 20, pp. 433–446.

[154] Skyrme, D. 2001. Capitalizing on Knowledge: From e-business to k-business? Butterworth-Heinemann, London.

[155] Smeaton, A., Lee, H., Colum, F., McGivney, S. 2006. Collaborative video searching on a tabletop. Multimedia Systems, 12(4), pp. 375—391, September 2006.

[156] Smith. G. 2004. Folksonomy: social classification. Atomiq. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html, August, 3, 2004.

[157] Smyth, B., Balfe, E. 2006. Anonymous personalization in collaborative web search. Information Retrieval 9 No.2, pp. 165-190.

[158] Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M., and Boydell, O. 2004. Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine. User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 14(5), pp. 383–423.

[159] Smyth, B., Briggs, P., Coyle, M. and O'Mahony, M. 2009. Google shared: a case-study in social search. In Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization (UMAP'09), pages 283 – 294.

[160] Spark, D. 2009. Real-Time Search and Discovery of the Social Web. Spark Media Solutions. http://www.sparkminute.com/?p=1261.

[161] Spence, P. R., Reddy, M. and Hall, R. 2005. A Survey of Collaborative Information Seeking of Academic Researchers. In GROUP'05, Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work, ACM Press, pp. 85-88.

[162] Surowiecki, J. 2004. The Wisdom of Crowds: Why the many are smarter than the few and how collective wisdom shapes Business, Economies, Societies and Nations. Little and Brown.

[163] Tapscott, D. and Williams, A. D. 2006. Wikinomics: How mass collaboration changes everything. Portfolio.

[164] Teevan, J., Adar, E., Jones, R., and Potts, M. 2007. Information Re-retrieval: Repeat queries in Yahoo's Logs. In Proc. SIGIR'07, ACM Press, pp. 151-158.

[165] Teevan, J., Dumais, S. T. and Horvitz, E. Personalizing search via automated analysis of interests and activities. 2005. In SIGIR'05, Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 449 - 456.

[166] Teevan, J., Ramage, D., and Morris, M. R. 2011. #TwitterSearch: a comparison of microblog search and web search. In Proc. WSDM '11. ACM Press, pp. 35-44.

[167] Tong, H., Faloutsos, C., and Pan, J. Y. 2008. Random walk with restart: fast solutions and applications. Knowledge and Information Systems, vol. 14, March 2008, pp. 327–346.

[168] Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., and Yahia, S.A. 2008. Efficient Computation of Diverse Query Results. Proc. The 24th International Conference on Data Engineering (ICDE 08), IEEE Computer Society, pp. 228-236.

[169] Victor, P., Cornelis, C., de Cock, M., & da Silva, P. P. 2009. Gradual trust and distrust in recommender systems. *Fuzzy Sets and Systems*, 160 (10), pp. 1367–1382.

[170] Vlachos, M., Meek, C., Vagena, Z., and Gunopulos, D. 2004. Identifying similarities, periodicities and bursts for online search queries. In: Proc. of ACM SIGMOD International Conference on Management of Data (SIGMOD 04), ACM Press, pp. 131-142.

[171] Voorhees, E. M., Harman, D. 1999. Overview of the 7th Text REtrieVal Conference (TREC-7). In E.M.Voorhees and D.K. Harman, editors, In *Proc. of the 7th Text RE-trieval Conference* (TREC-7), pp. 1-23.

[172] Wang, Q., Mayer-Schonberger, V. 2010. The Monetary Value of Virtual Goods: An Exploratory Study in MMORPGs. The 43rd Hawaii International Conference on System Sciences (HICSS), IEEE Computer Society, pp.1-11.

[173] Wang, X., and Zhai, C. 2007. Learn from Web Search Logs to Organize Search Results, In SIGIR'07, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 87-94.

[174] Wang, Y. & Mainwaring, S.D. 2008. "Human-Currency Interaction": Learning from Virtual Currency Use in China. In Proceedings of CHI 2008 (Florence, Italy, 5-10 April, 2008), ACM Press, pp. 25-28.

[175] Wasserman, S. and Faust, K. 1994. Social Network Analysis: Methods and Applications. Cambridge University Press.

[176] Watts, D. J. and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks, *Nature*, London, vol. 393, (June 1998), pp. 440-442.

[177] Wei, J., Bressan, S., Ooi, B.C. 2000. Mining term association rules for automatic global query expansion: methodology and preliminary results, in WISE'00, pp. 366–373.

[178] Wei, X. and Croft, W.B. 2006. LDA-based document models for ad-hoc retrieval. In SIGIR'06, ACM Press, pp. 178–185.

[179] Wellman, B. 1996. An Electronic Group Is Virtually a Social Network. Culture of the Internet. S. Kiesler. Hillsdale, NJ, Lawrence Erlbaum, pp. 179-208.

[180] Wen, J. R., Nie, J. Y., and Zhang, H. H. 2001. Clustering user queries of a search engine. Proc. The 10th international conference on World Wide Web (WWW 01), ACM Press, pp. 162–168.

[181] Wen, J. R., Nie, J. Y., and Zhang, H. J. 2001. Query clustering using content words and user feedback. In SIGIR'01, ACM Press, pp. 442–443.

[182] White, J. V., and Fournelle, C. G. 2005. Threat Detection for Improved Link Discovery. In Proceedings of International Conference on Intelligence Analysis.

[183] White, R. W., and Drucker, S. M. 2007. Investigating behavioral variability in web search. Proc. The 16th international conference on World Wide Web (WWW 07), ACM Press, pp. 21-30.

[184] Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., and Li, H. 2010. Context-aware ranking in web search. In SIGIR'10, ACM Press, pp. 451–458,

[185] Xie, Y. and O'Hallaron, D. 2002. Locality in search engine queries and its implications for caching. In INFOCOM'2002, Proceedings of the 21st Annual

Joint Conference of the IEEE Computer and Communications Societies, IEEE Computer Society, pp. 1238-1247.

[186] Xiong, L., and Liu, L. 2004. Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. IEEE Transactions on Knowledge and Data Engineering, 16, pp. 843–857.

[187] Xu, J. and Croft, W. B. 1996. Query expansion using local and global document analysis. In SIGIR'96, ACM Press, pp. 4–11.

[188] Xu, S., Bao, S., Fei, B., Su, Z., and Yu, Y. 2008. Exploring folksonomy for personalized search. In Proc. SIGIR '08. ACM Press, 155-162.

[189] Xue, G. R., Zeng, H. J., Chen, Z., Yu, Y., Ma, W. Y., Xi, W. S., and Fan, W. G. 2004. Optimizing web search using web click-through data. In CIKM'04, ACM Press, pp. 118–126.

[190] Yahoo! Media Relations: For Immediate release, new research finds that search plays a significant role with college students and new parents. http://docs.yahoo.com/docs/pr/release1267.html.

[191] Yamaguchi, H., An Analysis of Virtual Currencies in Online Games (September 1, 2004), Available at SSRN: http://ssrn.com/abstract=544422.

[192] Yanbe, Y. Jatowt, A. Nakamura, S. and Tanaka, K. 2008. Can social bookmarking enhance search in the web? In JCDL'08, Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, ACM Press, pp. 107-116.

[193] Yang, J., Ackerman, M. S. and Adamic, L. A. 2011. Virtual Gifts and Guanxi: Supporting SocialExchange in a Chinese Online Community. Proceedings of the international conference on Computer Supported Cooperative Work, Hangzhou, P. R. China, March 21-23, 2011, pp. 45-54.

[194] Yang, M., Chen, H., Zhao, B. Y., Dai, Y. and Zhang, Z. 2004. Deployment of a large-scale peer-to-peer social network. In WORLDS'04, Proceedings of the 1st Workshop on Real, Large Distributed Systems, December 2004.

[195] Yu, B., Singh, M. P. 2002. An evidential model of distributed reputation management. In Proc. of AAAMAS'02, pp. 294- 301.

[196] Zhai, C., Cohen, W., and Lafferty, J. 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. Proc. The 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR 03), ACM Press, pp. 10-17.

[197] Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., Ma, W-Y. 2005. Improving Web Search Results Using Affinity Graph. The 28th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR 05), ACM Press, pp. 504-511.

[198] Zhang, X., Li, Y. and Jewell, S. 2005. Design and evaluation of a prototype user interface supporting sharing of search knowledge in information retrieval. In: A. Grove (ed.), In ASIST'2005, Proceedings of the 68th Annual Meeting of the American Society for Information Science and Technology.

[199] Zhang, Z. & Jasimuddin, S. M. 2008. Pricing strategy of online Knowledge Market: the analysis of Google answers, International Journal of E-Business Research, 4, 1, pp. 55-68.

[200] Zhang, Z., and Nasraoui, O. 2006. Mining search engine query logs for query recommendation. In: Proc. The 15th International Conference on World Wide Web (WWW 06), ACM Press, pp. 1039-1040.

[201] Zhao, H., and Li, X. 2008. H-Trust: A Robust and Lightweight Group Reputation System for Peer-to-Peer Desktop Grid. In Proc. of the 28th IEEE ICDCS Workshops, pp. 235–240.

[202] Zhou, D., Bian, J., Zheng, S., Zha, H. and Giles C. Lee. 2008. Exploring Social Annotations for Information Retrieval. In WWW'08, Proceedings of the 17th World Wide Web Conference, ACM Press, pp. 715-724.

[203] Zhou, R., and Hwang K. 2007. Powertrust: A robust and scalable reputation system for trusted peer-to-peer computing. IEEE Transactions on Parallel Distributed Systems, 18, pp.460–473.

[204] Zhou, R., and Hwang, K. 2007. Gossip-based Reputation Aggregation in Unstructured P2P Networks. In Proc. IEEE International Parallel and Distributed Processing Symposium.

[205] Ziegler, C. and Golbeck, J. 2007. Investigating interactions of trust and interest similarity. Decision Support System. 43, 2, pp. 460-475.

[206] Ziegler, C-N., and Lausen, G. 2005. Propagation models for trust and distrust in social networks, Information Systems Frontiers 7 (4–5) (2005), pp. 337–358.

[207] Zubiaga, A., García-Plaza, A. P., Fresno, V. and Martínez, R. 2009. Content-based Clustering for Tag Cloud Visualization. In ASONAM'09, Proceedings of International Conference on Advances in Social Networks Analysis and Mining.

Publications Derived from this Research

1. Y. Mao, H. Shen, and C. Sun, Supporting exploratory information seeking by epistemology-based social search. In Proceedings of *the 15th ACM International Conference on Intelligent User Interfaces* (*IUI '10)*: 353-356.

2. Y. Mao, H. Shen, and C. Sun, EPISOSE: An Epistemology-based Social Search Framework for Exploratory Information Seeking. In Proceedings of *the 2nd IFIP TC 13 Human-Computer Interaction Symposium (HCIS 2010)*: 211–222.

3. Y. Mao, H. Shen, C. Sun. A Mutual Feedback Search Scheme on Real-time Web. *The 11th International Workshop on Collaborative Editing Systems (IWCES'11), Jointly with ACM CSCW 2011*.

4. Y. Mao, H. Shen, C. Sun. Looking for Non-existent Information: A Consumer-led Interactive Search Approach. In Proceedings of *the 25th British Computer Society Conference on Human-Computer Interaction (BCS HCI 2011)*.

5. Y. Mao, H. Shen, C. Sun. Google+Facebook: A Social-Network-Optimized Web Search Approach. *The 6th International Workshop on Ubiquitous and Collaborative Computing (iUBICOM'11), in conjunction with BCS HCI 2011*: 1-8.

6. Y. Mao, H. Shen, C. Sun. A Social-knowledge-directed Query Suggestion Approach for Exploratory Search. In Proceedings of *the 3rd IEEE International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC 2011)*:1-8.

7. Y. Mao, H. Shen, C. Sun. A Probability-based Query Expansion Approach to Informational Search Tasks. In Proceedings of *the 7th International Conference on Advanced Data Mining and Applications (ADMA 2011)*: 109–123.

8. Y. Mao, H. Shen, C. Sun. Diversification of Web Search Results through Social Interest Mining. In Proceedings of *the 45th Annual Hawaii International Conference on System Sciences (HICSS-45)*: 3581-3590.

9. Y. Mao, H. Shen, C. Sun. Distributed Trust Inference through Credit Flow in Online Social Networks. From Credit and Risk to Trust: Towards a Credit Flow Based Trust Model for Social Networks. In Proceedings of *the 2012 ACM SIGCHI International Conference of Supporting Group Work.* (Group 2012) .

10. Y. Mao, H. Shen, C. Sun. Online Silk Road: Nurturing Social Search through Knowledge Bartering. *Proceedings of the ACM 2013 conference on Computer Supported Cooperative Work* (CSCW 2013) (under review).

Appendix A

This appendix contains details about the classic information retrieval models introduced in Chapter 2.

The Boolean model is one of the earliest and simplest retrieval methods, which uses "exact matching" to match documents to a user query. More refined descendants of this model are still used by most libraries. In this model, keywords are logically combined with the Boolean operators AND, OR, and NOT, therefore a document is judged as relevant or irrelevant based on whether keywords are present or absent in the document. There is no concept of a "partial match" between documents and queries, and the inability to identify partial matches can lead to poor performance [11].

The vector space model of information retrieval is a very successful statistical method proposed by Salton [142]. It generates weighted term vectors for each document in the collection, and for the user query. Retrieval is based on the similarity between the query vector and document vectors, and the output documents are ranked according to this similarity. The similarity is based on the occurrence frequencies of the keywords in the query and in the documents.

In this model, documents and queries are represented as vectors in a t-dimensional space. Each dimension corresponds to a separate term. The dimensionality of the vector is the number of distinct terms in the vocabulary. If term i occurs in document d_j, the term is

associated with a non-zero weight $w_{i,j}$ in the document vector. As more frequent terms in a document are more important than others, i.e., more indicative of the topic, a well-known combined term importance indicator is TF-IDF weighting. Let $f_{i,j}$ be the frequency of term t_i in document d_j, the normalized term frequency (*tf*) across the entire corpus is given by:

$$tf_{i,j} = \frac{f_{i,j}}{\sum_i f_{i,j}}$$

Let N be the total number of documents in the corpus, and df_i be the document frequency of term t_i (the number of documents containing term t_i). Then idf_i is the inverse document frequency of term t_i:

$$idf_i = \log \frac{N}{df_i}$$

The *idf* value is an indication of a term's discrimination power. The logarithm is used to dampen the effect relative to *tf*. The most well-known heuristic TF-IDF for computing document relevance is given by:

$$w_{i,j} = tf_{i,j} \times idf_i = tf_{i,j} \times \log \frac{N}{df_i}$$

A term occurring frequently in the document but rarely in the rest of the collection is given high weight.

The query is also transformed into a vector. The similarity between vectors for document d_j and query q (or score of d_j for q) can be computed as:

$$score(d_j, q) = \sum_{i=1}^{t} w_{i,j} \times w_{i,q}$$

where $w_{i,j}$ is the weight of term t_i in document d_j and $w_{i,q}$ is the weight of term t_i in the query q.

The vector space model addresses some of the problems of the Boolean model, and remains quite competitive as a popular retrieval model. However, the drawbacks of this model are its computational expense and poor scalability.

The probabilistic model is based on the Probability Ranking Principle [17], which states that documents in a collection should be ranked based on their probability of relevance to the query. Given a user query, the probabilistic model tries to estimate the probability that the user will find the relevant documents. The model assumes that this probability of relevance depends only on the query and the document representations.

Denoting the probability of relevance for document d_j to query q by $P(R \mid D)$, and $P(\overline{R} \mid D)$ is the probability that the document is irrelevant. Using Bayes rule and assuming the terms occur in a document independent of each other, the probabilistic ranking is computed as:

$$score(d_j, q) = \frac{P(R \mid d_j)}{P(\overline{R} \mid d_j)} = \frac{P(d_j \mid R)P(R)}{P(d_j \mid \overline{R})P(\overline{R})} \propto \frac{P(d_j \mid R)}{P(d_j \mid \overline{R})} = \frac{\prod_i P(t_i \mid R) \times \prod_i P(t_i \mid \overline{R})}{\prod_i P(\bar{t}_i \mid R) \times \prod_i P(\bar{t}_i \mid \overline{R})}$$

Since the ranking criteria are monotonic under log-odds transformation, we can take logarithms and rewrite the ranking formula using only the values for terms present in a document:

$$score(d_j, q) \propto \log \frac{\prod_i P(t_i \mid R) \times \prod_i P(t_i \mid \overline{R})}{\prod_i P(\bar{t}_i \mid R) \times \prod_i P(\bar{t}_i \mid \overline{R})} \propto \sum_i \log \frac{P(t_i \mid R) \times (1 - P(t_i \mid \overline{R}))}{P(t_i \mid \overline{R}) \times (1 - P(t_i \mid R))}$$

Note that if we think of $\log \frac{P(t_i \mid R) \times (1 - P(t_i \mid \overline{R}))}{P(t_i \mid \overline{R}) \times (1 - P(t_i \mid R))}$ as the weight of term t_i in document d_j, this formulation becomes very similar to the similarity formulation in the vector space model with query terms assigned a unit weight.

The probabilistic model takes into account that there is uncertainty in the representation of the information need and the documents. However, the simplistic assumption that words are independent of each other is not realistic. For instance, in this thesis the most likely word to follow “probabilistic” is “model”, therefore it is not reasonable to assume that these two words are independent.

In recent years, a new family of probabilistic models called (statistical) language models have shown good performance compared with other traditional models. Before they were applied to information retrieval [127], language models had been used in automatic speech recognition systems for many years. The first use of language model for IR, which is generally referred to as the query-likelihood retrieval model and was first proposed by Ponte & Croft [127], is to estimate a language model for each document and rank documents by the likelihood of generating the submitted query. Berger and Lafferty [18] proposed the statistical translation model, which determines the relevance of a document

to a query by estimating the probability that the query would have been generated as a translation of that document. Documents are then ranked according to these probabilities. The relevance model proposed in [90] is a description of a user's information need, and the ideal relevance model for a given query run on a specified document collection would be constructed. Each document relevant to the user's query then becomes a sample from the underlying relevance model. Lafferty and Zhai [89] introduced a risk minimization framework based on Bayesian decision theory. In this framework, queries and documents are modeled using statistical language models, user preferences are modeled through loss functions, and retrieval is cast as a risk minimization problem.

The recent application of statistical language models to information retrieval has proven to be immensely successful. However, to capture a user's real intent and context of an information retrieval process, language models must be integrated with other techniques to support more advanced retrieval tasks.

Appendix B

This appendix contains detailed discussions of the comparison between current web search frameworks and the EPISOSE framework presented in Chapter 3.

EPISOSE is uniquely distinguished from current web search frameworks listed in Figure B.1(a)-(d). Figure B.1(a) shows the conventional method of web search with a search engine. In these systems, the *Search Engine* is an algorithm-based search engine such as Google. Search knowledge cannot be shared since there is no knowledge generation and storage mechanism. Peer to peer search engines can support each peer to publish its local documents (or local index), but the purpose is to build efficient topic-specific search engines rather than supporting general EIS through knowledge sharing. Figure B.1 (b) shows the method of web search with QA websites, e.g. asking for help in forums or Yahoo! Answers, which is also known as the man-powered search engines. Users can share knowledge through those social web sites, but knowledge sharing is not effective as shared knowledge still has to be retrieved by conventional search engines. Figure B.1 (c) shows the method of web search with social annotations, such as Mahalo, Yoople, Scour[23], and Wikia Search[24], which allows people to publish their attitudes toward certain search results through voting or editing. But annotations are not as rich in content as search epistemologies (annotations are only an integral part of epistemologies) and

[23] http://www.scour.com
[24] http://search.wikia.com

sharing of annotations is not effective as they have also to be retrieved by search engines. Figure B.1 (d) shows the method of web search with real-time collaborative tools such as SearchTogether [115], which supports search and sharing of Web pages with others through communication facilities. Searched Web pages cannot be shared to people who are not invited or online when the search process is ongoing. Moreover, it shares only Web pages rather than a package of knowledge about a specific topic.

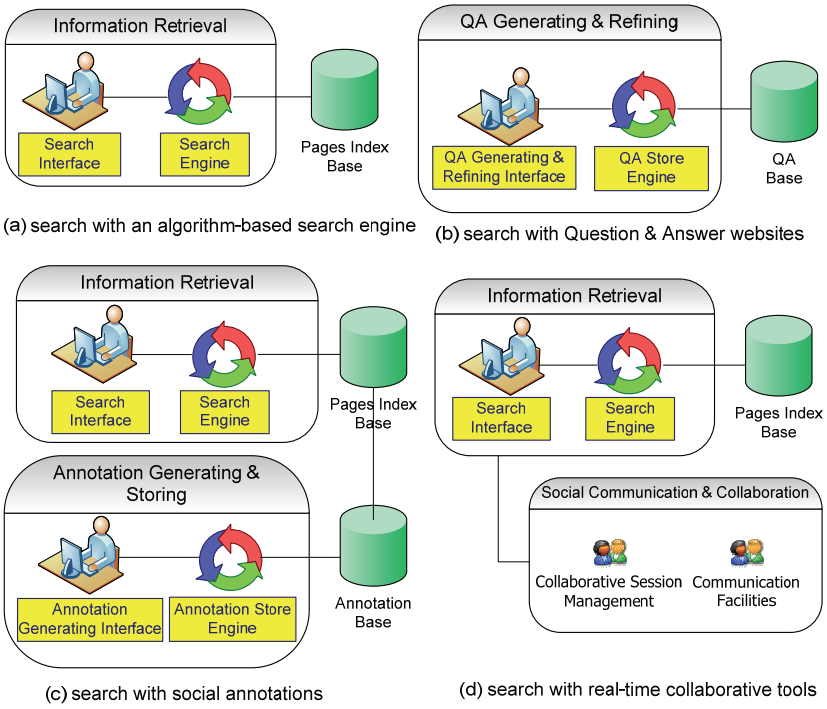

Figure B.1 A comparison of EPISOSE with other Web search frameworks

The primary goal of a social search system for exploratory information seeking is to utilize existing successful searches. Many people have searched for the same topic, for example, wedding planning, but they couldn't re-use previous successful searches because they were not shared or fell into oblivion. Yet it's not true that people are unwilling to share the knowledge gained from their searches. The problem is that they couldn't find convenient and effective ways to share it. As a matter of fact, many people

write their knowledge of wedding planning in blogs or discussion forums after effortful web searches. However, only a small group of people in a specific period of time may benefit from such kind of knowledge sharing, while others still need to spend tremendous time repeating the process of searching for the same topic.

The EPISOSE framework addresses this problem as follows. First, it provides a social search platform and interfaces for a wide range of prosumers to share their search epistemologies directly. Second, search epistemology is based not only on a rich set of objective results returned by a search engine, but also on a prosumer's subjective judgment and intimate knowledge. It is far beyond just a set of search keywords and a list of linear results. Third, searching and sharing are seamlessly integrated in the framework so that prosumers can enrich their knowledge about a search topic and improve their search skills by learning from their peers. It is worth pointing out that an epistemology-based social search system is self-reinforcing in the sense that epistemologies tend to be more relevant and accurate as they are refined by more prosumers.

Appendix C

This appendix illustrates the details of the epistemology-based social search in the *Baijia* prototype system described in Chapter 3.

C.1 Epistemology Search and Generation

Take a search of wedding planning as an exploratory information seeking example. A user (AnonID=19913) begins with the keyword "wedding planning". After the user inputs "wedding planning" in the search interface (Figure 3.5), the system will return a linear list of results through the common API of a search engine (currently the Google AJAX Search API). If previous search epistemologies exist, a re-ranked results list will be returned. The details of these results show who have contributed to the search epistemologies and how do other users evaluate these epistemologies, which will help the user judge their relevance.

After browsing a few pages of the results, the user regards the page from "www.weddingsolutions.com" as the best page about the solutions of wedding planning and drags the item from "search results" to "best results". The user can also add some personal comments on this result, e.g., "This is a good website about wedding planning". The interfaces for search and sharing are presented in one page so that users can easily drag and drop items and read and write comments without any popup windows.

Having gained the basic knowledge about wedding planning, the user may continue the search for “wedding receptions” if she/he considers having a wedding reception (AnonID=884092), or “wedding dresses” if she/he wants to get some good suggestions about the dressing in the wedding ceremony (AnonID=5761104). She/he can choose “share epistemology” to publish the epistemology after completing the whole search process.

She/he may entitle the epistemology as “Wedding Planning Introduction” and classify it into a “successful search” or “partially successful search” category. She/he can even supplement the search with her/his own epistemology: “All those preparations are not enough and I think there should be…”.

When another user also starts exploratory information seeking about wedding planning with the keyword “wedding planning”, she/he will get a list of related epistemologies from others (Figure 3.6). The relevant search epistemologies can be collected by filtering according to the similarity to the query. For a user who knows little about wedding planning, what has been searched by others is a good starting point for her/him. Hence, the user can browse the list and dig out the details of those in which she/he is interested, e.g. “Wedding dresses for a bride”, if she/he (e.g. AnonID=16852248) wants to know more about wedding dresses. The user can also get help from others’ epistemologies if she/he is not sure about what she/he is searching for. For example, the user (AnonID=12199341) has only a vague idea about “how can I write the wedding invitations”. After browsing the list, she/he finds that many others are looking for “wedding invitation wording”, which effectively suggests her/him how to start the search.

Furthermore, a user can learn search skills from others' epistemologies. For example, when a user (AnonID=2200929) wants to search information about a band playing music in the wedding ceremony and she/he uses the keyword "wedding band", she/he would be puzzled since all results returned by a search engine in the first page are all about the wedding rings. If a previous successful search epistemology named "Music Wedding Band" has been generated in the epistemology repository, the user can easily find relevant information from the related epistemologies list. Meanwhile, she/he knows how to formulate keywords and to filter out unwanted ones. Such knowledge will probably help accelerate her/his search process.

The user can continue her/his search on the basis of these searches, add new good results and commentaries or remove expired or faulty information. If a page is dynamic, she/he may add annotations or modify the link with parameters to ensure it will not expire. The user can also evaluate others' epistemologies, e.g., give positive or negative feedback. All search epistemologies will be accumulated in the epistemology repository for further utilization.

C.2 Epistemology Search Engine & Epistemology Repository

The epistemology search engine derives the relevance between search epistemology and a search query. For example, if the "www.weddingsolutions.com" page of wedding has appeared as the best result in both the search epistemologies of "Wedding Planning Guide" and "Traditional Wedding Planning", we are likely to draw the conclusion that it is relevant to the query about "wedding planning". Besides, some methods such as

feature extraction and clustering will be employed to accurately discover more relevant search epistemologies to certain search goals.

For conflicting search epistemologies, the reputation and the expertise of users can be used to help make the judgment. One's reputation will increase/decrease when one's published search epistemologies receive positive/negative evaluation by others. This will discourage users from publishing misleading or irresponsible search epistemologies for malicious purposes or undeserved reward points. The epistemologies published by users with better reputation will be assigned heavier weights.

Since all search epistemologies are stored in the epistemology repository, its volume grows fast with time. Therefore, it is important to build suitable indices for search epistemologies in order to improve the retrieval efficiency and accuracy. Moreover, with the epistemology repository, we can also support semantic epistemology retrieval by building various ontologies. For example, the "band" would be modeled with the meaning of "ring" for an ontology about the domain of wedding, but with the meaning of "instrumentalists" for another ontology about the domain of music. As such, our system could process users' requests more accurately based on the context of their queries.

Appendix D

This appendix contains detailed technical motivations for the RTU-LDA model proposed in Chapter 4.

It is non-trivial to build a topic model for epistemology generation, as it must discover semantically related queries in an EIS process and rank results based on their relevance to the topic of the process, through the statistical analysis of many users' EIS activities.

First, the query space is very sparse as different users often use different queries for the same or a similar search goal, which is known as the "vocabulary problem" [47]. Therefore if the click-through data were modeled as a "bag of words", where the queries were the "words", it would be difficult to discover the relationships between the queries due to the few recurrences of queries. However, different users may use the same query terms in their queries to describe the same or a similar search goal and these terms are usually combined with different other terms in different queries. For example, one user generated two queries in sequence: "how to speed up windows xp" and "security system virus", and another user generated two different queries in sequence: "windows system speed" and "anti virus update". It would be difficult to discover the relationships between these queries as they are regarded as four different "words" in their entireties, albeit these queries are intrinsically related (to the topic about making Windows XP faster). In contrast, it would be easier to discover that the query terms of "windows", "speed", and "virus" are all related to the same or a similar search goal (for speeding up Windows) by

statistical analysis of the co-occurrence of these terms. Therefore, our approach is to discover query relevance by inferring how the query terms are related to the general goals of EIS tasks because the term space dimension is much smaller than the query space dimension.

Second, the query clustering techniques based on the K-means algorithm [63] do not specifically address the polysemy problem in EIS because the algorithm places each query into one and only one cluster without considering the intrinsic semantic coherence among query terms (e.g., only based on syntactic term matches). A term may have diverse meanings, e.g., "windows" could either refer to a computer operating system or components of a house. Therefore, a better solution is to consider the diverse meanings of each term in a query and discover related queries whose terms are logically and semantically coherent. Our approach is to build a topic model that can discover the latent semantic topics and the probabilities of the query terms being associated with each of these topics. Based on the model, the queries are related if they have similar probability distributions over topics decided by the probabilities of their terms being associated with the same topic. For example, query q_1= "how to speed up windows" is more related to q_2="anti virus update" than q_3="house with three windows", because q_1 and q_2 both have high probabilities of being associated to a latent topic on "computer" and low probabilities of being associated to a topic on "estate", while q_3 has a high probability of being associated with the topic on "estate" but a low probability of being associated with the topic on "computer".

Third, analysis of query logs based on search sessions suffers from the "topic shift" problem [120]. A user may generate "noisy queries" (i.e., queries that are not related to

the goal of the current search session), which may impede mining query relevance. Topic models such as pLSI (Probabilistic Latent Semantic Index) [68] and LDA relax the assumption made in the models with mixture of unigrams (e.g., LSI [41]) that each document is generated from a single topic. Therefore, each user's click-through data can be treated as one document consisting of a mixture of different topics. Our topic model is a significant extension of LDA, where large-scale click-through data from a vast number of users is treated as a collection of documents. The topic mixture is drawn from a conjugate Dirichlet prior [26], which is a sharp contrast to pLSI, where topic mixture is limited to each individual document.

Last, applying LDA directly onto the query terms may not necessarily discover the latent semantic topics that are relevant to the search goals as these query terms themselves may not precisely describe the search goals in the first place. We address this issue by incorporating social annotations (e.g., Delicious[25] that allows users to share and annotate their bookmarks with relevant tags) into our topic model. The rationale is twofold. First, the tags added to a Web page by the users who have viewed the page must contain their intimate knowledge in understanding and classifying the content. Second, the tags added to a page tend to be more normalized (i.e., using commonly acceptable and understandable terms) than queries used to retrieve the page. Relevant topics can be derived from the query terms and the social annotations of the retrieved Web pages to infer users' actual goals of their EIS tasks in the training process. After that, more precise query terms could be discovered to expand the initially imprecise queries with the aid of

[25] http://delicious.com

the derived topics. For example, if the “computer” tag were added to most URLs clicked by the users who generated queries related to “how to speed up windows xp”, the query term of “windows” would have a high probability of being associated with the latent topic on “computer”, and conversely a low probability of being associated with the topic on “estate”.

Appendix E

This appendix provides the query suggestion evaluation for the social-interest-directed approach presented in Chapter 5.

The query suggestion results of our approach are compared with the following methods in our experiments:

- Query Clustering (QC) method [10], which is a clustering process over data extracted from the query log to identify groups of semantically similar queries;
- Forward Random Walk (RW) method [38], which is a query-independent random walk proceeds to its stationary distribution on the query-URL bipartite graph;
- Hitting Time (HT) method [108], which is a query-dependent parameter-free random walk on the query-URL bipartite graph using hitting time.

It is difficult to evaluate the quality of query suggestion due to the scarcity of data that can be examined publicly. Further, there is no ground truth for "correct" queries to suggest because different judgments could be made even by human experts. Therefore, we conduct both automatic evaluations based on the *Google Directory* or a commercial search engine (i.e., *Google*), and manual evaluations by a panel of four human experts to assess the relevancy and diversity of query suggestion respectively.

E.1 Relevancy Assessment

First, we evaluate the relevancy of suggested queries. We made both objective and subjective comparison between different methods to give a fair assessment.

- Automatic Evaluation:

  We evaluate the relevancy of suggested queries by utilizing a similar method used in [173]. Specifically, we measure the relevancy of two queries based on the similarity between their corresponding categories provided by the *Google Directory* instead of the *Open Directory Project*. When a user types a query in *Google Directory*, besides site matches, we can also find category matches in the form of paths between directories. Moreover, these categories are ordered by relevancy. For example, the query "Palm OS" would provide the hierarchical category "Computers > Systems > Handhelds", while one of the results for "BlackBerry" would be "Computers > Systems > Handhelds > Smartphones > BlackBerry". Therefore, to measure how related the two queries are, we can use a notion of similarity between the corresponding categories provided by the search results of *Google Directory*. In particular, we measure the similarity between two categories C_i and C_j as the length of their longest common prefix $P(C_i, C_j)$ divided by the length of the longest path between C_i and C_j. More precisely, the similarity is defined as:

$$sim(C_i, C_j) = \frac{\left|P(C_i, C_j)\right|}{\max(\left|C_i\right|, \left|C_j\right|)}$$

where $|C_i|$ denotes the length of a path. For example, the similarity between the above two queries is 3/5 since they share the path "Computers > Systems > Handhelds" and the longest one is made of five directories. We evaluate the similarity between two queries by measuring their similarity between the aggregated categories, among the top 5 answers provided by *Google Directory*.

For the metric of this evaluation, we adopt the precision at rank k to measure the relevancy of the top k results of the suggested list with respect to a given query q_j, which is defined as:

$$P@k = \frac{\sum_{i=1}^{k} sim(q_i, q_j)}{k}$$

where $sim(q_i, q_j)$ means the similarity between q_i and q_j. We report the precision from $P@1$ to $P@10$, and take the average over all the 100 distinct queries in our experiments.

- Manual Evaluation:

  For the manual evaluation, we ask all the experts to rate the results of query suggestion in a 6-point scale (0 to 5) that measures the relevancy between the suggested queries and the testing queries, where 0 means "completely irrelevant" and 5 indicates "perfectly relevant".

E.2 Diversity Assessment

Second, the diversity of suggested queries is evaluated. We automatically compute the diversity score with the help of *Google*, and manually make diversity judgments with human assessment efforts.

- Automatic Evaluation:

  To determine the quality of diversity between queries literally is not easy. Therefore in our experiment, we first compute a measure of diversity of queries in the resulting set based on the differences between their top ranked search results provided by a search engine. This is done by taking each sampled query and each suggested query, and issuing the queries to *Google*. Specifically, given two queries q_i and q_j, Uq_in is the set of top n (n = 10 in our case) URLs in the search results of query q_i, we compute the difference of their search results between q_i and q_j by:

$$diff(q_i, q_j) = 1 - \frac{\sum rel(Uq_in, Uq_jn)}{n}$$

  where $rel(u_{q_in}, u_{q_jn})$ measures the relevancy between two URLs Uq_in and Uq_jn among the top n search results of query q_i and q_j. Given that we are taking the top k suggested queries, this generates a maximum of $k \times n$ URLs. If we use the union of the queries which lead to clicking URL Uq_in to represent this URL, the relevancy $rel(Uq_in, Uq_jn)$ is calculated by Cosine similarity using the representations of URLs Uq_in and Uq_jn. Intuitively, if q_i and q_j represent two

entirely different concepts or topics, then Uq_in and Uq_jn probably have no intersection, and the URLs in the set Uq_in are dissimilar with those in the set Uq_jn. Therefore a large value of $diff(q_i, q_j)$ means queries q_i and q_j are diverse. For instance, the diversity value between queries "amazon coupons" and "amazon web services" is greater than the value between queries "amazon coupons" and "amazon discount code" in our dataset.

Then for an input query q, we use the metric diversity at rank k to measure the diversity of the top k results of its suggested list, which is defined as:

$$D@k = \frac{\sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} diff(q_i, q_j)}{k(k-1)}$$

We report the average diversity values from $D@2$ to $D@6$ of the suggestion result sets for all of the 100 distinct queries in the testing dataset.

- Manual Evaluation:

  Similar to the relevancy assessment, we ask all the experts to rate the results of the query suggestion for epistemology-based EIS in a 6-point scale (0 to 5) that measures the diversity between the testing queries and the testing queries, where 0 means "scarcely diverse" and 5 indicates "highly diverse".

E.3 Results

Now we consider the problem whether our proposed social interest directed (*SID*) approach could generate queries that are more diverse but still relevant using the query-URL bipartite graph for query suggestion.

The results of the relevancy assessment are presented in Figure E.1 and Figure E.2. Figure E.1 illustrates the precisions of four methods based on the automatic evaluation from *P*@1 to *P*@10. In general, we can see that the performance of our method is better than the baseline methods. It also demonstrates the effectiveness of incorporating social knowledge to capture the relevancy between queries. After looking into the details, one important observation is that the improvements of our method over the baselines are increased for larger *k* (of the evaluation metric *P@k*). This is probably because the social knowledge can boost the relevant long tail queries that have low initial scores.

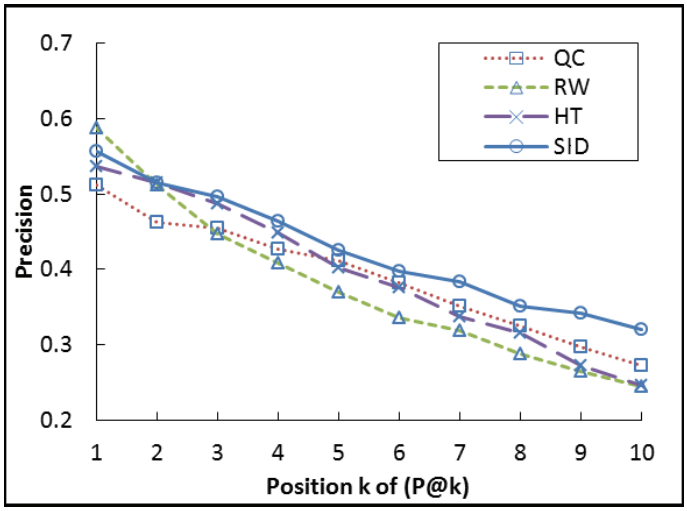

Figure E.1 Average relevancy of query suggestion: *P@k* of four methods

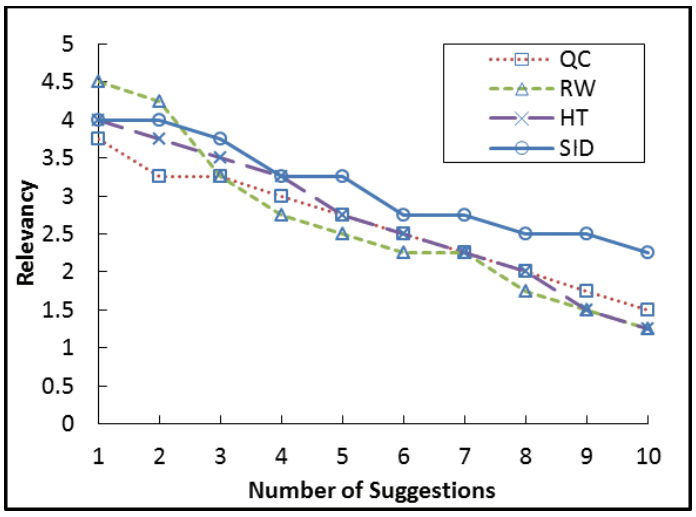

Figure E.2 Relevancy ratings of four methods given by experts

Figure E.2 shows the average values of manual evaluation results. Similarly, the queries suggested by our method retrieve better epistemologies as compared to *QC*, *RW* and *HT* methods when measuring the results by human experts. Moreover, the improvement is more significant than the measurement based on *Google Directory*, since *Google Directory* is not complete enough to discover the semantic similarity between queries, while human experts can better understand the latent relationship between the query and epistemologies which are not related directly.

Figure E.3 and Figure E.4 shows the experimental results for the diversity assessment, and all these evaluations are performed on our method and the three baselines. Figure E.3 presents the results of the automatic evaluation from *D*@2 to *D*@6 for these methods. The diversity of *QC* is the lowest one in the four methods, as *QC* focuses on suggesting queries according to the similarity with the given query. *RW* gets better diversity than *QC*, but it sacrifices the relevancy (See Figure E.2) considerably. *HT* leverages the hitting time from candidate queries to the given query as their ranking scores, and thus boosts

the long tail queries for suggestion and increases the diversity. Therefore, it obtains a higher diversity than both *QC* and *RW*.

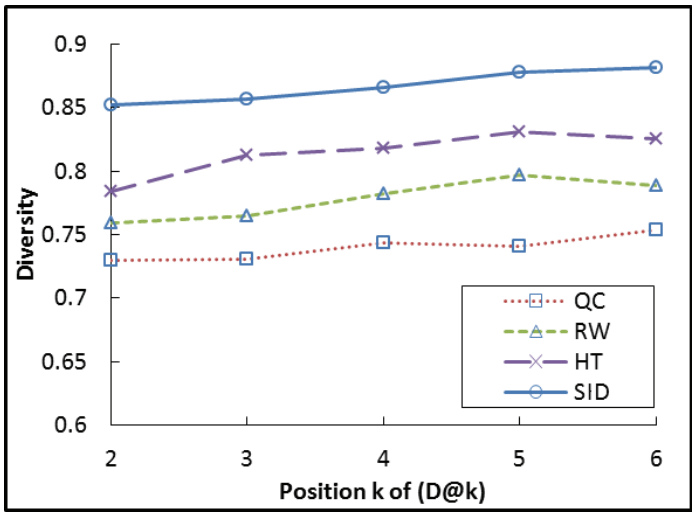

Figure E.3 Average diversity of query suggestion: *D@k* of four methods

However, among all these four methods, *SID* obtains the highest diversity as it explicitly addresses the diversity problem by introducing the social knowledge space to replace the query space. Further, the experimental results of the manual evaluation illustrated in Figure E.4 also demonstrate the effectiveness of our proposed diversified query suggestion algorithm.

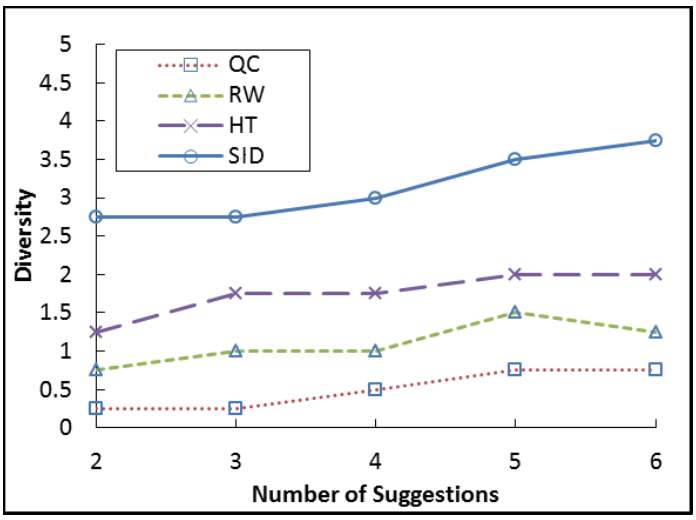

Figure E.4 Diversity ratings of four methods given by experts

Appendix F

This appendix contains the detailed reciprocation of social networks and social search in the IPOD approach described in Chapter 6.

F.1 Building Social Networks from Social Search Activities

In a search process, the prosumer usually needs to formulate a set of queries $sp = \{q_1, q_2 \ldots q_n\}$, and the epistemology of this search *Epi(sp)* is defined as:

$$Epi(sp) = Epi(q_1) \cdots \oplus Epi(q_i) \cdots \oplus Epi(q_n) \oplus Epi(extra),$$

where $Epi(q_i)$ is the epistemology for q_i ($1 \leq i \leq n$), and *Epi(extra)* is the epistemology for related information that is not acquired through these queries, such as information from authoritative websites, and '⊕' is the operator to construct the epistemology for a search process out of those for constituent queries. For each $Epi(q_i)$, the definition is based on the prosumer's interaction with the system. Such as pages selected by the prosumer: $\{p_1, p_2 \ldots p_m\}$, and the prosumer's ranking and comments on the pages.

Therefore there are two spaces in the epistemology-based social search: epistemology space and prosumer space. Each prosumer might participate in several epistemologies and each epistemology might be contributed by several prosumers. Prosumers are able to modify contents in the epistemology space by the epistemic operations such as:

Add(q_i): add a new query q_i to the epistemology;

Remove(q_i): remove the query q_i;

Tag(epi_i): append a new tag for epistemology epi_i;

Annotate(q_i): annotate the entry q_i;

Select(p_j): select the page p_j from an epistemology;

Rank(p_j): rank page p_j;

Comment(p_j): add comments on page p_j.

Figure F.1 shows the two spaces in the epistemology-based social search.

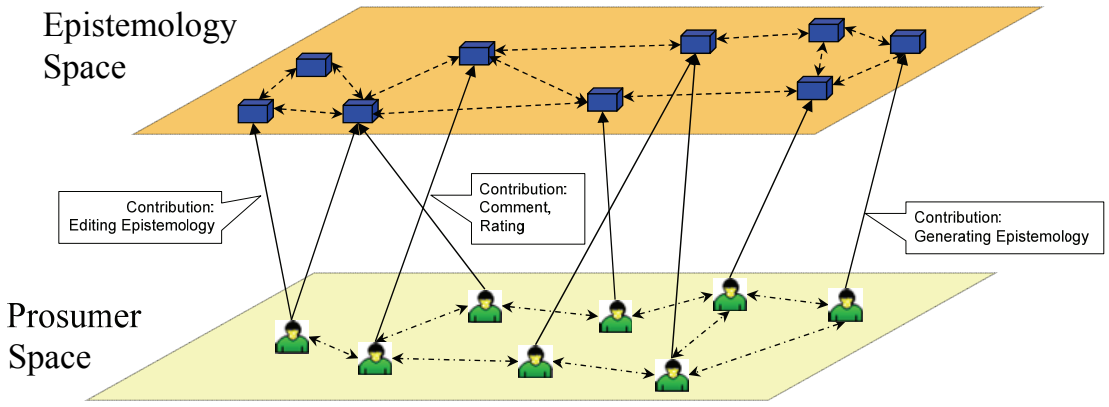

Figure F.1 Epistemology-based social search spaces

The connection between epistemologies can be derived from the content of each epistemology. The more two epistemologies are relevant, the shorter the distance between them is. The distance is defined based on the summation of all epistemic concepts, each of which is contributed by prosumer *i*, and the match is based on the similarity between all elements of the concept in two epistemologies:

$$Dist(Epi_1, Epi_2) = \sum_i \sum_k (w_1 \times D(q_i, q_k) + w_2 \times D(tags_i, tags_k)$$
$$+ w_3 \times D(comments_i, comments_k) + w_4 \times D(p_i, p_k)), q_i, p_i \in Epi_1, q_k, p_k \in Epi_2$$

where w_i is the weight assigned to an element of the epistemology according to its importance, e.g., an element such as a comment or a page with a higher prosumer ranking will be assigned a heavier weight.

The similarity between two elements can be measured by various methods. In our approach, it is measured with the Kullback-Liebler divergence (KL-divergence) [89] between two language models [127]. For example, if a query q_1 and a query q_2 is generated by a generative model P_{q1} of epistemology Epi_1 and P_{q2} of epistemology Epi_2 respectively, their KL-divergence is defined as:

$$D(q_1 \| q_2) = \sum_w P_{q1}(w) \log \frac{P_{q1}(w)}{P_{q2}(w)}$$

where $P_{q1}(w)$ is the probability of generating word w by the language model for query (q_1), $P_{q2}(w)$ is the probability of generating word w by the language model for query (q_2).

Smoothing techniques are usually used for language models due to the data sparseness problem. The smoothing technique we adopted in our approach is Jelinek-Mercer (JM), therefore given a document D, the probability of generating word w is:

$$P(w \mid D) = \lambda P_{ML}(w \mid D) + (1 - \lambda) P_{ML}(w \mid Coll), \lambda \in [0,1]$$

where $P_{ML}(w \mid D)$ is the maximum likelihood estimator for a word w given a document D, $P_{ML}(w \mid Coll)$ is the maximum likelihood estimator for a word w in the collection

language model *Coll*, and λ is a parameter controlling the amount of mass distribution assigned to the document and collection.

Now we consider correlating prosumers with epistemologies. The social network in the social search community is a weighted graph $G = (U, E)$, where each node represents a prosumer and each edge $e = (u_1, u_2)$ is the correlation between prosumers u_1 and u_2. For each epistemology *Epi* contributed by one or more prosumers from the community $U = \{u_1, u_2, \ldots, u_m\}$, we have a set of keywords $Epi = \{k_1, k_2, \ldots, k_n\}$.

We first model the contribution of a prosumer u to an epistemology *Epi* using a monotone aggregate function g over the individual relevance for each keyword k in *Epi*:

$$Con(u \mid Epi) = g(rev(u \mid Epi, k_1), ..., rev(u \mid Epi, k_n)),$$

where $rev(u \mid Epi, k)$ is the relevance of prosumer u and epistemology *Epi* for a keyword k in *Epi*. The aggregation function g used in this article is a summation: $g = \sum_{k_i \in Epi} rev(u \mid Epi, k_i)$. We use a TF-IDF scoring function [143] to measure the relevance, which amounts to a simplified form of BM25, as follows:

$$rev(u \mid Epi, k) = \frac{(p+1)\, freq(u \mid Epi, k)}{p + freq(u \mid Epi, k)} \times idf(k),$$

where p is an application dependent parameter, $freq(u \mid epi, k)$ is the overall term frequency of u given the epistemology *epi* and keyword k, i.e., the number of times k was quoted by prosumer u, and $idf(k)$ is the inverse document frequency for keyword k, which is defined in fairly standard manner as follows:

$$idf(k) = \log \frac{|Epi| - |\{u \mid Quoted(u, Epi, k)\}| + 0.5}{|\{u \mid Quoted(u, Epi, k)\}| + 0.5}$$

Then the correlation between two prosumers u_1 and u_2 can be obtained from the epistemologies they contributed $\{E1_1, E1_2, ..., E1_m\}$ and $\{E2_1, E2_2, ..., E2_n\}$ respectively:

$$Corr(u_1, u_2) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|Con(u_1, E1_i) - Con(u_2, E1_j)|}{Con(u_1, E1_i) + Con(u_2, E1_j)} Dist(E1_i, E2_j) \quad (1)$$

The rationale behind formula (1) is that if two prosumers both made major contribution to some highly related epistemologies, they might have same or similar interest so that we can correlate them each other. For example, if a prosumer has contributed a lot to an epistemology about "World Cup", and another prosumer has deeply involved in an epistemology about "Messy", as these two epistemologies has many overlapped keywords, e.g., "goal", "champion", we can deduce that the two prosumers are both soccer fans and there is a great opportunity that they can make friends online, because they can talk with and learn from each other when searching for common topics on the Web.

F.2 Exploring Social Networks for Social Search

We have built social networks of likeminded prosumers who appear to have similar preferences in the social search community. The main purpose of exploring the social networks is to locate prosumers that can be helpful for a consumer in her/his future search processes: predict potential information providers for her/him, or recommend to

her/him the epistemologies contributed by Top-*N* trustworthy prosumers and that she/he would like the most.

The connection between trust and user similarity has been established by Ziegler and Golbeck [205]. They used experiments to demonstrate that there exists a significant correlation between the similarity of users and the trust expressed by them; the more similar two people are, the greater the trust between them.

In addition, user reputation is a powerful method of identifying high-quality providers over time and has been adopted in some social web applications, where reputation is based on feedback on items that user has created in the past, and serves as a signal of quality as well as an incentive to improve quality.

In our approach, epistemologies have been rated and commented by other prosumers, and therefore the system will re-rank the epistemologies in the repository dynamically. The ranking of an epistemology is based on all received scores (one to five stars) for all pages in the epistemology, and the reputation (honest or fraudulent) and expertise (newcomer or skilled) levels of each contributor and commenter.

Based on epistemology-mediated social networks and the user reputation, our approach utilizes the user model to generate a cluster of most similar and trustworthy prosumers in the social network for a consumer and then to identify the Top-*N* prosumers in the cluster that have gained highest reputation and can act as providers or advisers to that consumer.

The *k*-means clustering algorithm is applied in the social network analysis. For initializing *k*-means, the *k* "means" $\mathbf{m}_1^{(1)},\ldots,\mathbf{m}_k^{(1)}$ are initialized with prosumers randomly

selected from the social network. The algorithm *k*-means proceeds by alternating between two steps [99]:

Assignment step: Each prosumer is assigned to the cluster with the closest mean (i.e. partition the users according to the Voronoi diagram generated by the means).

$$S_i^{(t)} = \{\mathbf{x}_j : \left\|\mathbf{x}_j - \mathbf{m}_i^{(t)}\right\| \le \left\|\mathbf{x}_j - \mathbf{m}_{i^*}^{(t)}\right\| \text{ for all } i^* = 1,...,k\}$$

where $\left\|\mathbf{x}_j - \mathbf{m}_i^{(t)}\right\|$ is the correlation between users calculated by formula (1).

Update step: Calculate the new means to be the centroid of the prosumers in the cluster.

$$\mathbf{m}_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j$$

The algorithm is deemed to have converged when the assignments no longer change.

Appendix G

This appendix introduces the power flow study in electrical systems as the background for the credit-flow-based trust model presented in Chapter 7.

In a modern commercial electrical system, it is necessary to analyze the system performance with given network configurations. The analysis in normal steady-state operation is called a power flow study, which calculates the voltage drop on each feeder, the voltage at each bus, and the power flow in all branch and feeder circuits. We first introduce power flow calculation and power balance equations, where power flow variables are to be determined. We then explain the Newton-Raphson method expressed in the rectangular form for solving the power flow equations.

G.1 Electrical Power System

An electrical power system is a network of electrical components that supply, transmit and consume electrical power. All power systems have three major types of components: *Generation*, supplying electrical power; *Load*, consuming electrical power; and *Transmission*/*Distribution*, transmitting electrical power from generation to load. Each power system has one or more sources of power and the alternating current (AC) power is typically supplied by a turbo generator. Loads in power systems range from industrial machinery to household appliances. The sources are connected to the loads via conductors such as transmission lines. For instance, Figure G.1 shows a simple power

system, consisting of 5 buses (equal to nodes), 7 transmission lines, 2 generators, and 3 loads.

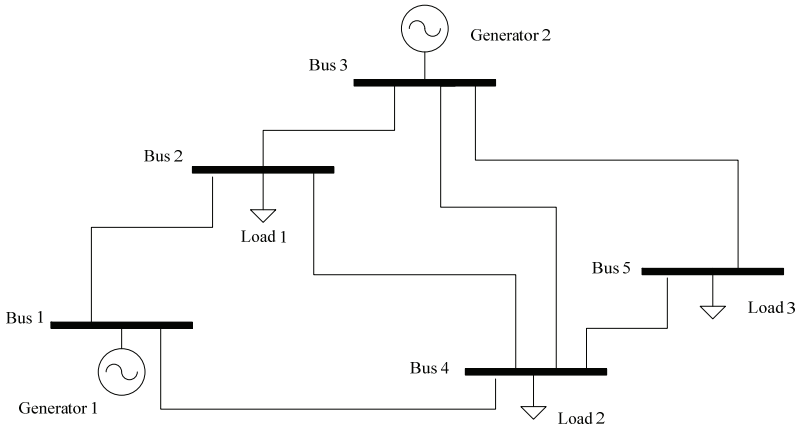

Figure G.1 Example of a simple power system

Because of the significant advantages of AC over direct current (DC) in transmission, today's electrical power systems are dominated by AC power. If a load contains inductance or capacitance that can store electrical energy in every AC cycle, the electrical energy will periodically return into the AC power supply, flowing back and forth across the conductors and consequently producing an extra current. That energy is *reactive power*, which is opposite to the *active* (or *real*) *power* supplied by a generator for the load to do real work, e.g., lighting a bulb.

If the structure of an electrical network and the impedances of all components within the network are known, we can use node voltage analysis methods to solve all the currents in the network with Kirchhoff's circuit law. Given the vector $\boldsymbol{V}$ of voltages at all nodes and the vector $\boldsymbol{I}$ of currents injected at each node bus, the basic node equation for a power system is:

$$I = Y_{bus} V \qquad (1)$$

where $\boldsymbol{Y_{bus}}$ is the nodal admittance of the buses in the system. $\boldsymbol{Y_{bus}}$ is formulated with the following two rules if any one of the buses is chose as the reference: 1) the admittance of elements connected between bus k and the reference is added to entry (k, k) of the admittance matrix; and 2) the admittance of elements connected between buses i and k is added to entries (i, i) and (k, k) of the admittance matrix, while the negative admittance is added to entries (i, k) and (k, i) of the admittance matrix.

However, because real loads are specified in terms of active and reactive powers not currents, it is not possible to directly use the node equations. Therefore, we use the power balance equations, which are applicable to both real power P and reactive power Q. The relationship between real and reactive powers supplied to the system from a bus and the current injected into the system at that bus is:

$$S_i = V_i I_i^* = P_i + jQ_i, \qquad I_i = \frac{P_i - jQ_i}{V_i^*}$$

where V_i is the per-unit voltage at the bus; I_i^* is the complex conjugate of the per-unit current injected at the bus; P_i and Q_i are per-unit real and reactive powers.

Substituting I_i with formula (1) yields:

$$P_i - jQ_i = V_i^* \sum_{k=1}^{N} Y_{ik} V_k \qquad (2)$$

However, because not enough variables are usually known a priori to solve Equation (2), e.g., the real power and voltage are known but reactive power is unknown. Power flow study is to solve this algebraic non-linear equation with iterative techniques.

G.2 Newton-Raphson Method for Power Flow Study

Power flow study begins with identifying known and unknown variables, which depend on the type of bus: load bus - a bus without any generators connected to it, generator bus - a bus with at least one generator connected to it, or slack bus - only one such bus in an electrical system. Due to energy loss in an electrical network, real and reactive power cannot be known at all buses. Therefore, the slack bus can provide the necessary power to maintain the power balance in the network. The voltage magnitude $|V|$ and phase angle θ (normally set to zero degree) are known for the slack bus, the real power P_D and reactive power Q_D are known for load buses, and the real generated power P_G and the voltage magnitude $|V|$ are known for generator buses.

Several methods exist for solving the non-linear equations in power flow study and a widely used one is the Newton-Raphson method. It uses sequential linearization to transform the original non-linear problem into a sequence of linear problems and solutions to these linear problems approach the solution to the original non-linear problem.

We first re-formulate the power balance equations with real coefficients, where G_{ik} is the real part of entry (i, k) in the bus admittance matrix $\boldsymbol{Y_{bus}}$, B_{ik} is the imaginary part of entry (i, k) in the bus admittance matrix $\boldsymbol{Y_{bus}}$, and θ_{ik} is the voltage angle difference between the ith and kth buses.

$$P_i - jQ_i = \sum_{k=1}^{n} |V_i||V_k| e^{j\theta_{jk}} (G_{ik} + jB_{jk}) = \sum_{k=1}^{n} |V_i||V_k| (\cos\theta_{ik} + j\sin\theta_{ik})(G_{ik} + jB_{jk})$$

Separating the real and imaginary parts,

$$P_i = \sum_{k=1}^{N} |V_i||V_k|(G_{ik}\cos\theta_{ik} + B_{ik}\sin\theta_{ik}) \qquad (3)$$

$$Q_i = \sum_{k=1}^{N} |V_i||V_k|(G_{ik}\sin\theta_{ik} - B_{ik}\cos\theta_{ik}) \qquad (4)$$

The Newton-Raphson method begins with an estimation of the unknown variables, which is called initial guess. Subsequently, correction to these estimations is done by expanding Equations (3) & (4) in Taylor's series about the initial estimate and neglecting the higher order terms resulting in an iterative calculation process:

$$\begin{bmatrix} \Delta P \\ \Delta Q \end{bmatrix} = \mathbf{J}\begin{bmatrix} \Delta\theta \\ \Delta|V| \end{bmatrix} = \begin{bmatrix} \frac{\partial\Delta P}{\partial\theta} & \frac{\partial\Delta P}{\partial|V|} \\ \frac{\partial\Delta Q}{\partial\theta} & \frac{\partial\Delta Q}{\partial|V|} \end{bmatrix}\begin{bmatrix} \Delta\theta \\ \Delta|V| \end{bmatrix} \qquad (5)$$

where ΔP and ΔQ are the differences between the scheduled and calculated values, known as the power residuals:

$$\Delta P_i^{(j)} = \sum_{k=1}^{N} |V_i||V_k|(G_{ik}\cos\beta_{ik} + B_{ik}\sin\beta_{ik}) - P_i^{(j)} \qquad (6)$$

$$\Delta Q_i^{(j)} = \sum_{k=1}^{N} |V_i||V_k|(G_{ik}\sin\beta_{ik} - B_{ik}\cos\beta_{ik}) - Q_i^{(j)} \qquad (7)$$

and **J** is the Jacobian matrix specifying the linearized relationship between the small changes in real and reactive power ΔP and ΔQ and the small changes in voltage magnitude $\Delta|V|$ and phase angles $\Delta\theta$.

After that, the linearized system of equations is solved to determine the next guess ($j+1$) of voltage magnitude and phase angles based on the following recursions:

$$\theta^{(j+1)} = \theta^{(j)} + \Delta\theta\,,\ |V|^{(j+1)} = |V|^{(j)} + \Delta|V|$$

The iterative process continues until a stopping condition is met. A common stopping condition is that the residuals ΔP and ΔQ are less than the specified accuracy.

In summary, the Newton-Raphson method solves the power flow problem through the following steps:

1) *Make an initial guess of all unknown voltage magnitudes and angles. It is common to use a flat voltage start in which all voltage magnitudes are set to 1 and angles are set to 0.*

2) *Use the most recent voltage magnitude and phase angle values to solve the power balance equations.*

3) *Calculate the entries in the Jacobian matrix to linearize the system around the most recent voltage magnitude and phase angle values.*

4) *Solve the linear simultaneous equation (5) directly through optimally-ordered triangle factorization and Gaussian elimination.*

5) *Compute the new voltage magnitudes and phase angles from equations (6) and (7).*

6) *Terminate if a stopping condition is met, otherwise go to step 2.*