Rheinische Friedrich-Wilhelms-Universität Bonn
Institut für Informatik III

Proposal for a Diploma Thesis in Computer Science

Social Information Retrieval

Sebastian Marius Kirsch*
Advisor: Prof. Dr. A. B. Cremers

4th May 2005

Abstract

Despite recent efforts in personalization and collaboration, information search on the internet is still mostly anonymous. It does not take the user's long-term interests or his environment into account.

In this diploma thesis, we want to research whether the inclusion of additional information about a user's social environment can lead to a significant improval in search effectiveness.

We propose to model the search domain as an associative network and to evaluate two different techniques for information retrieval in such an enviroment. We will compare these techniques to standard vector space retrieval for evaluation purposes.

1 Introduction

1.1 Information Retrieval and the Social Realm

The goal of information retrieval is facilitating a user's access to information that is relevant to his information needs. According to Baeza-Yates and Ribeiro-Neto (1999), an information retrieval system 'should provide the user with easy access to the information in which he is interested.' Earlier definitions took a narrower and more technical view on the purpose of a retrieval system, for example Lancaster (1968): 'An information retrieval system does not inform (i. e. change the knowledge of) the user

*skirsch@moebius.inka.de

on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.', or Frakes and Baeza-Yates (1992): 'An I[nformation] R[etrieval] system matches user *queries* – formal statements of information needs – to documents stored in a database.'

An information retrieval system must first determine the exact nature of the user's information needs, then select a subset of documents that help him satisfy his information need, and finally rank the selected documents according to which documents are most likely to provide a satisfactory answer.

Wilson (1981) notes that both the user's information needs and his strategies for satisfying them are influenced by the socio-cultural environment, since they arise in social situations. Wenger (1996) introduced the idea of the 'community of practice': the notion that person can satisfy his information needs more efficiently if he is embedded in a community of practitioners with similar interests and problems. Indeed, before the advent of modern information retrieval systems, most information needs were satisfied by social means: by asking friends and acquaintances, by going to the library and asking the librarian for help, or by enquiring at specialized agencies.

Although the amount of information available in automated retrieval systems is far greater than can be acquired from other people, information that comes from immediate contacts is usually preferable to information obtained from anonymous sources: Since the provider is known, it is easier to assess the quality of the information. Here, quality has several different aspects; the first and foremost is factual accuracy. But there are also secondary aspects, for example the provider's subjective evaluation, the ability to further discuss the topic with the provider, and obtain references to other relevant pieces of information. Only when one's immediate contacts are not able to satisfy the information need or more in-depth information about a topic is required, one turns to secondary sources – equipped with the information acquired by asking within the community.

Information retrieval meets the social realm at another, more subtle point: Information is also produced in social situations. Few authors work in a social vacuum. Participation in the community and active exchange with like-minded persons fosters information production and improves the quality of the work.

With the increasing use of electronic communications media, viz. the Internet, social ties and the structure of the social network become tractable. We have identified a number of systems (see section 5) where data about social ties between users is available, in addition to similarity data or references between documents and information about authorship. In such a setting, incorporating social information into the retrieval process is an obvious next step: Since both information usage and information production occur in social environments, both are influenced by the user's and the author's social network. Knowledge of the network affects all parts of the information retrieval

problem.

We therefore define social information retrieval as the incorporation of information about social networks and relationships into the information retrieval process.

1.2 Paving the way: Personalized and Collaborative Information Retrieval

In conventional information retrieval systems, all of the user's information needs are embodied in a query: a short string of key words or a question. Further indicators of the user's general information needs are not taken into account, such as his previous searches or his web sites of interest. Indeed, a query with one or two keywords is much too short to contain a complete picture of a user's needs. A search engine is therefore susceptible to a form of tyranny of the majority: It can only display those sites that will be relevant to the majority of its users, but not to the actual user who submitted a query.

Personalization seeks to solve this problem by keeping a record of the user's previous activity and using it to attune the results to his profile. Implementations of personalized search exist, but are not yet in widespread use; examples are Amazon's a9.com[1] and Eurekster[2], which are implemented as a central service, or SearchPad (Bharat, 2000), a client application.

A collaborative element can be added by comparing and combining the profiles of different users. This approach is popular in information filtering systems such as the GroupLens system (Konstan et al., 1997) for filtering Usenet posts. It has also been used in information retrieval systems, for example in the aforementioned Eurekster system, or the experimental I-Spy[3] search engine (Freyne and Smyth, 2004).

Personalization strategies and collaborative retrieval attack the problem of determining a user's information needs from different angles. Personalization aim to infer a more detailed view of the information needs based on past usage, whereas collaborative ranking acknowledges that the information seeker is part of a community of like-minded individuals.

1.3 Social retrieval and the World Wide Web

Much if not most of the current research in information retrieval is focused on searching the World Wide Web, a topic that at the same time presents inherent obstacles (due to its size and its lack of structure) and great promises (due to the amount of information that is publicly available.) Extracting the most relevant pages from 8 billion web

[1] http://www.a9.com, last visit on 2005/04/11.
[2] http://eurekster.com, last visit on 2005/04/11.
[3] http://ispy.ucd.ie/, last visit on 2005/04/15.

pages[4] is a daunting task, especially if all information about the desired results is condensed to one or two keywords. (Silverstein et al. (1999) give an average length of 2.35 keywords for their analysis of AltaVista query logs.) During the evolution of internet search engines, it quickly became apparent that this problem cannot be solved by relying only on automatic evaluation of a web page's content, but needs some sort of human assessment of a page's relevance.

Early attempts to build a manual index of web pages, selected by human editors, (so-called 'web catalogues') were largely unsuccessful – because of the sheer size of the web and the limited manpower of the companies. Most major web portals still provide some kind of directory, for example the Google Directory[5] or the Yahoo! Directory[6], or use data from the Open Directory Project[7]. However, the focus for navigating the web has been on automated information retrieval, not manual indexes, for several years.

Recent efforts in collaborative projects have shown that it is possible to garner a large, active user community, in the tens of thousands or even millions of users, within a few months. Projects such as Wikipedia[8] show that large undertakings purely on the basis of volunteer labour are possible. (In this sense, PageRank (Page et al., 1999) is also a collaborative effort in information retrieval and ranking, since it uses link information published on millions of web pages.)

These examples motivate a vision for the future of web search that is not dominated by centralistic efforts of single companies, providing us with results derived from a global view of the web. One may envision a service that provides each user with results that are tailored to his individual information needs, and that derives its results by collaborating with other users, sharing information and relevance assessments. Such a tool would be an ideal application for social information retrieval, since it combines the social network with the wealth of information available on the World Wide Web.

One tool that aims to implement this vision is discussed in the next section.

1.4 ISKODOR: Congenial Web Search

ISKODOR[9] is an experimental system developed at the University of Bonn that seeks to implement all three approaches for enhancing information retrieval presented in the last section. The stated goal of the project is the implementation of 'congenial web search' (Gnasa et al., 2004) – meaning a user-centred approach where search quality is constantly evaluated through explicit feedback.

[4] According to their front page, the Google search engine indexes 8,058,044,651 as of 2005/04/14.
[5] http://www.google.com/dirhp, last visit on 2005/05/08.
[6] http://dir.yahoo.com/, last visit on 2005/05/08.
[7] http://www.dmoz.org/, last visit on 2005/05/08.
[8] http://www.wikipedia.org/, last visit on 2005/04/08.
[9] ISKODOR is an acronym for 'Is Sharing Knowledge Online a Dream Or Reality?'

The functional prototype of ISKODOR employs a peer-to-peer architecture in order to share search results with other users. Thus, a single point of failure or bottlenecks are avoided. The user's faith in the service is strenghened, as he himself controls which information is stored and disseminated about him.

ISKODOR implements personalized ranking matrices; collaborative information retrieval is implemented in the form of peer groups, which are used to limit the scope of a search (Gnasa et al., 2003).

An ISKODOR peer can keep track of the quality of the results provided by its peers and re-rank results according to the peer that supplied it. This 'peer relevance' judgement leads to a network of trusted peers that produce the most relevant results.

Social search techniques can be applied in this network of trusted peers, to provide better search results and find peers that are well versed in a specific topic. Thus, social information retrieval can be used to improve web search effectiveness.

2 Related work

As noted by Romano et al. (1999), the connection between information retrieval and social processes has not been extensively researched to date. Even though information seeking has long been recognized as a social process (Wilson, 1981, 1994), few projects support social interaction in the information retrieval process or exploit this fact to achieve better performance.

The integration of social interaction into the information retrieval process is part of the domain of computer-supported collaborative work (CSCW). Masinter and Ostrom (1993) pioneered the field of collaborative information retrieval by allowing the Gopher system to be queried from the context of a MUD, thereby allowing real-time interaction and discussion about search results.[10] Romano et al. (1999) expanded this approach by designing a web-based system for collaborative information search, doing away with the need for real-time interaction by allowing users to annotate web pages; however, he noted that the collaborative features of the system were rarely used.

A different approach to the problem of finding information through social processes is expert location: finding a person in a corporate intranet or on the web who is likely to have the needed expertise. The ExpertFinder system (Mattox et al., 1999; Maybury et al., 2001) relied on information in company communications (eg. newsletters) and personal homepages, ranking experts by the frequency of mentions. Kautz et al. (1997a,b) tried to solve the problem by building chains of referrals, using information gathered from the Web. D'Amore (2004) uses the notion of activity spaces in order to

[10] Gopher (Anklesaria et al., 1993) is a 'distributed document search and retrieval protocol' that was largely eclipsed by the World Wide Web in the 1990s. MUD stands for 'Multi-User Dungeon' and is a class of multi-user, text-based, real-time chat systems designed for role-playing games.

generate aggregate scores to rank experts in a certain field. A more detailed analysis of the domain is found in (Yimam-Seid and Kobsa, 2003).

Aggregating judgements from a community of people for the purpose of ranking documents was first applied in collaborative filtering; it is also being integrated into search engines. The aforementioned I-Spy search engine (Freyne and Smyth, 2004) requires users to join a community and re-ranks search results according to the combined usage data from the community. This solves the problem of anonymity, since the usage data cannot be traced back to an individual user. One user can only be part of one community at a time; requiring the user to change the community as the subject matter of his search changes.

Incorporating social information into search processes creates the need for metrics that assess the strength of social ties and the standing of a member in the social network. Guha et al. (2004) studied a trust metric based on spreading trust and distrust through the network and concluded that small numbers of explicit opinions allowed a reliable computation of trust measures. Yamamoto et al. (2004) proposed PageRank as a reputation model for social networks and showed a way of calculating it in a peer-to-peer system, based on the messages exchanged between peers as part of their normal operation.

Finally, Yu and Singh (2003) studied searching on the basis of referrals in a multi-agent network, providing a method of limiting and directing search processes in a peer-to-peer network, based on social information.

3 Proposed approach

We propose to model the social information retrieval task as a search in an associative network. An associative network is a graph of information items, with unlabeled, weighted, directed or undirected edges ('associations') between nodes. We will use three kinds of nodes: Person nodes, signifying information produducers, content nodes, and query nodes. (Query nodes are special nodes representing keyword queries; they are transient in nature, and associations with other nodes in the network are computed on the fly.) The associations and weighting schemes will depend on the application domain. (See figure 1.)

The network can be queried for a node of any type, or for a combination of nodes; the search returns a list of nodes that are believed to be most relevant to the original nodes.

For search in the associative network, we intend to implement and evaluate two search techniques, both of which are essentially local in nature and can feasibly be implemented in a distributed manner. A local algorithm ensures scalability even for large networks. As one of the proposed applications of our techniques is a peer-to-peer

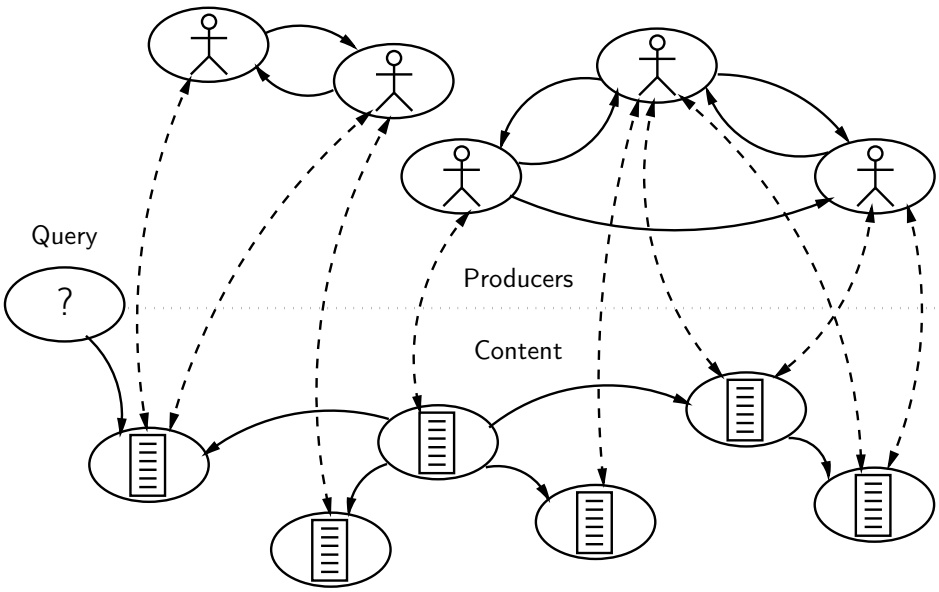

Figure 1: An associative network models the relationship between authors as well as between their works.

network, the algorithm should be suitable for a distributed implementation. We will compare our approaches to a conventional vector-space search method.

3.1 Spreading activation search

Spreading activation energy is a concept stemming from connectionist models of cognition (Anderson, 1983); it was popularized in the context of semantic networks. Spreading activation works by disseminating 'activation energy' from one primary node to secondary nodes via links between the nodes.

The application of spreading activation search for text retrieval was studied by Salton and Buckley (1988), who found its performance to be comparable to vector-space methods. Pirolli et al. (1996) used spreading activation to unify content-based and link-based information for searching the World Wide Web. Recent works (Ceglowski et al., 2003) re-evaluated the technique in the light of current developments, in particular spectral methods such as latent semantic indexing. An overview of spreading activation in information retrieval is found in (Crestani, 1997).

We think that spreading activation is a useful model for searching in social networks, as it is capable of transgressing nodes and search beyond similarities at the first level.

As a first experiment, we analyzed the coauthorship network of the SIGIR corpus. (See section 7 for a description of this corpus.) This coauthorship network contains 213 connected components; the largest contains 211 nodes and the smallest contain 2 nodes. The average component size is 4.73, the median component size is 3. The largest connected component is depicted in figure 2. We augmented this network by

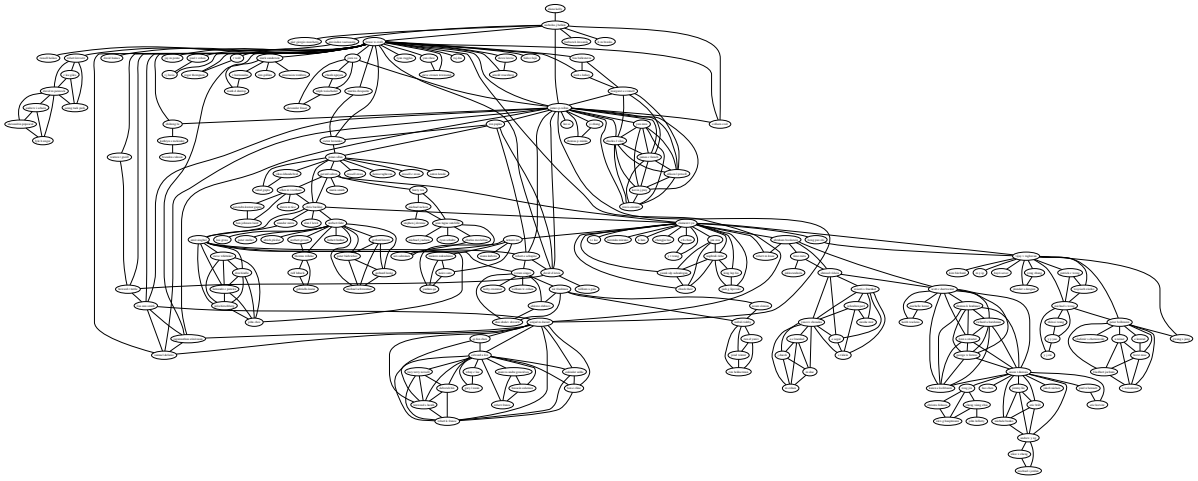

Figure 2: The largest connected component in the coauthorship network of the SIGIR corpus.

adding nodes for all published papers, and links between papers and their respective authors, as well as links between individual papers, based on the similarity of the abstracts.

A spreading activation search for an author with a dense coauthorship network, in our example we used Bruce Croft, shows a very shallow result that provides little more than the immediate coauthorships. (See figure 3) If we search for an author that is very much at the periphery of the SIGIR (B. John Ommen, who published only once in the SIGIR proceedings) shows a more differentiated picture (figure 4): We do not only find Oommen and his co-author, but also papers with a similar content, as well as their authors.

3.2 Maximum flow approach

Maximum flow techniques also use a network model of information. They were introduced as a method to solve the community identification problem in the web context (Flake et al., 2000, 2002, 2004). We believe that they are also a viable method for the 'neighbourhood' of a node or a set of nodes in an associative network, and hence can be used for search in such a network.

Maximum flow techniques are similar to spreading activation in that certain measure, 'flow', is distributed through the network via links. The amount of flow that may be distributed through a link is limited by the link's capacity; it is further constrained by the law of *flow conservation*, which states that the amount of flow that enters a node through incoming links must leave it again through outgoing links. Nodes that violate flow conservation are called 'sources' (if more flow leaves the node than enters it) or 'sinks' (if the reverse is true.)

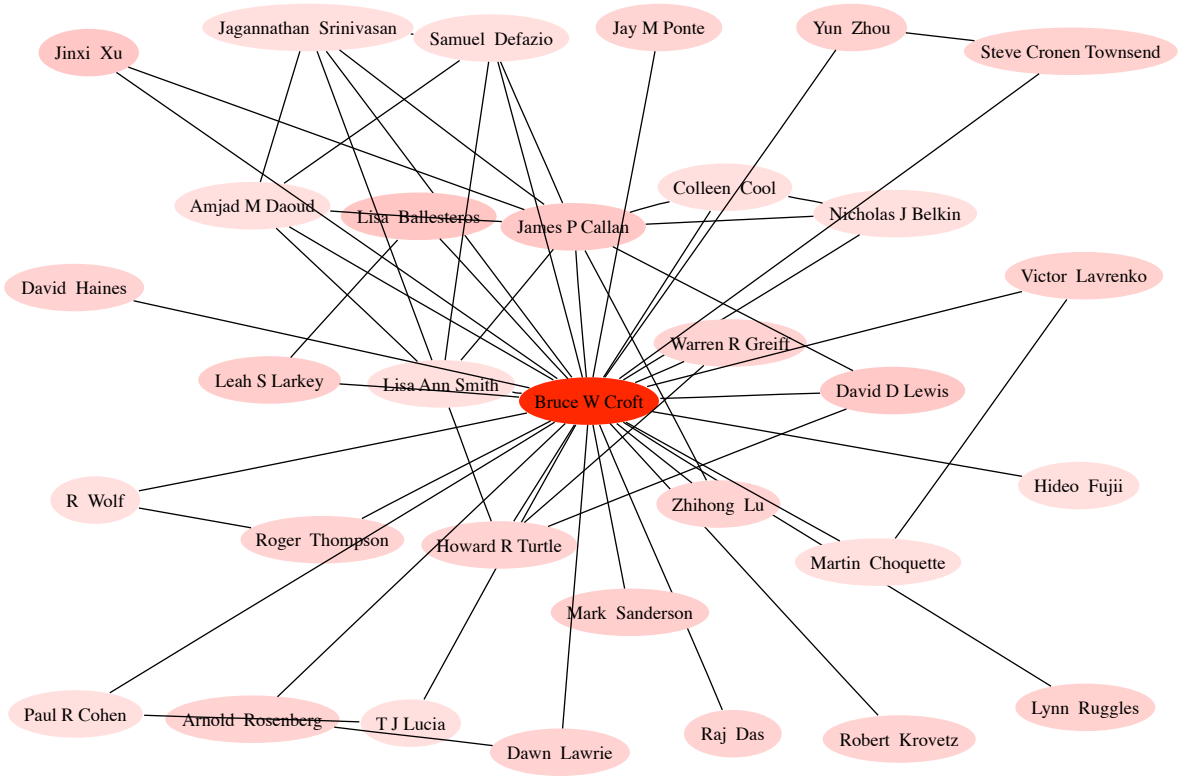

Figure 3: Spreading activation search for Bruce W. Croft in the SIGIR corpus. (Document nodes are hidden for better clarity in this figure.)

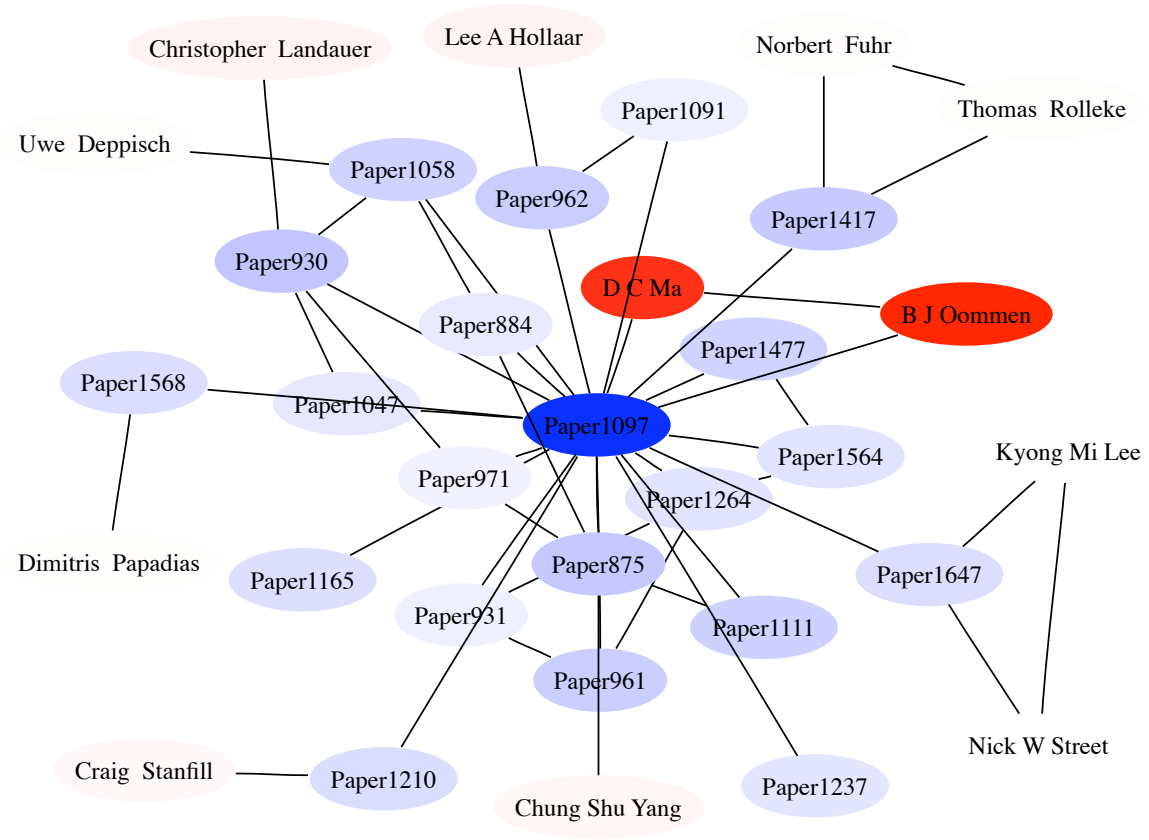

Figure 4: Spreading activation search for B. John Oommen in the SIGIR corpus. Oommen only published a single paper in the SIGIR proceedings; therefore, links to other members have to go through similarities between published papers.

Maximum flow search connects all activated nodes to an artificial node (the source) by links with infinite capacity; all original nodes are connected to an artificial sink by links with a fixed, small capacity. (These connections do not need to be stored explicitly, making maximum flow search essentially a local method.) The maximum flow between source and sink is then computed, returning all nodes through which flow is actually distributed.

3.3 Baseline method: Vector space search

In order to compare the performance of the previously mentioned techniques to common information retrieval techniques, we intend to implement a conventional vector-space search with appropriate weighting. (See (Frakes and Baeza-Yates, 1992; Baeza-Yates and Ribeiro-Neto, 1999).)

One crucial aspect of social information retrieval is the ability to search beyond immediate similarities and transgress nodes in the network. Since vector space search uses only similarity values, a simple keyword query would only yield documents containing the keywords. In order to select documents according to authorship and social relations of the author, a form of query expansion is necessary in this case. We will use pseudo relevance feedback (Xu and Croft, 1996) for query expansion.

4 Applications and benefits of social information retrieval

We expect that the proposed techniques will prove beneficial in several applications. We plan to formally evaluate only the first item.

4.1 Improved precision and quality of results for information retrieval

The basic premise of social information retrieval is that the inclusion of social information will lead to better search results. A retrieval system that contains author information and information about social networks can differentiate between authors at the fringe of the community and authors with good standing and reputation in the community, and use this information to assess document quality.

As noted previously, modern web retrieval systems already use collaborative evaluatons of document quality in the form of PageRank. PageRank as a global measure cannot be used to determine the relevance of a document *as regards the current topic*; the proposed local measures can take into account both the topic at hand and the importance of a documents in relation to it.

4.2 Personalized, collaborative search

Currently there is a lack of tools that allow a combination of collaborative and personalized search – producing search results tailored to a particular user, using information gathered from his community. Most efforts in collaborative search are based on using contributions from a fixed community of users, treating all community members as equal. This effectively replaced the lamented tyranny of the majority with a 'tyranny of the community', without catering to an individual user's needs.

Social information retrieval can provide this combination, since it does not operate on a fixed community of which a user is a member. Instead, every user with his social ties forms a community with himself at the center. Recommendations from other users are ranked according to their position in the social network, as regards the information seeker.

4.3 Support for nascent virtual communities

In nascent virtual communities, the community structure (as expressed through relationships between the members) is not very distinct. Through our hybrid approach that searches both the network of social relationships and the network of produced content, we facilitate the evolution of a community: Community ties can be formed between members with similar interests that did not interact previously, which will in turn augment the social network.

4.4 Improved scalability for distributed, social information retrieval

All modern information retrieval systems, except the most trivial ones, are distributed systems. Peer-to-peer applications and grid computing approaches aim to implement fast, scalable information retrieval systems without the tight coupling usually required for distributed systems; an example of a pure distributed information retrieval system is the Yacy[11] system. As individual peers in a peer-to-peer network can usually be associated with a single user, these systems provide a fertile ground for collaborative and social techniques.

The techniques in section 3 produce a subgraph of the associative network, which can be used to limit the scope of a search. If search algorithms in a distributed system can be limited to a subgraph of the network, their performance can be improved.

[11] http://www.yacy.net/yacy/, last visit on 2005/04/04.

5 Examples of social networks

We have identified a number of examples for social networks where the proposed approaches may be useful. The following section enumerates some of them, describes their nature and their use in the evaluation of our techniques.

In order to employ the proposed techniques, we require a social collaboration environment where the authorship of a piece of information can be readily determined, and where social ties between authors can be inferred.

5.1 ISKODOR

(See section 1.4.)

5.2 Scientific community

Social network analysis in the scientific community has a long tradition. Through the use of bibliometric measures such as co-citation coupling and bibliographic, the network structure of scientific publications and the publications they cite can be assessed..

A famous anecdotal application of network analysis in the natural sciences is a person's Erdős number[12]: The minimum length of a path in the co-authorship network between the Hungarian mathematician Paul Erdős and a given person.

Network analysis in the scientific community is usually conducted on the basis of publications in well-known journals or conference proceedings, as well as the cited publications. These documents usually do not capture the full extent of social relationships between authors, since much communication occurs via secondary channels, such as email. The observable content is of very high quality.

A number of databases of scientific publications exist, for example MathSciNet[13], PubMed[14] and CiteSeer[15]. Some databases, most notably CiteSeer, support download of records via the Open Archive Initiative Protocol for Metadata Harvesting[16].

A corpus with data from 25 years of SIGIR proceedings, stemming from work on (Smeaton et al., 2002) and enhanced locally, is available for evaluation.

[12] http://www.oakland.edu/enp/, last visit on 2005/03/08.
[13] http://www.ams.org/mathscinet, last visit on 2005/03/08.
[14] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed, last visit on 2005/03/08.
[15] http://citeseer.ist.psu.edu/, last visit on 2005/03/08.
[16] http://www.openarchives.org/OAI/openarchivesprotocol.html, last visit on 2005/05/02.

5.3 Wikis

Wikis are a form of collaborative authoring environment that is characterized by the fact that every user can add, edit, and delete content at will. The first wiki was WikiWikiWeb[17], launched by Ward Cunningham in 1995 as a supplement to the Portland Pattern Repository, a web site about software design patterns. A number of software packages and similar projects followed; the largest wiki is purported to be Wikipedia[18], an online encyclopedia that employs the wiki principles.

Wikis usually have a flat structure, with one designated entry page that links to other pages; some use fixed number of categories. Most wikis keep a revision history that allows changes to be linked to individual users. Direct interaction between users usually occurs on the user's home page.

The quality of published content varies wildly; some wikis contain nothing more than a few quickly written ideas, others, like Wikipedia, aim for publication-quality content.

The backend software and database dumps for Wikipedia are available for download; this is enticing for evaluation purposes. However, the current database dumps are very large, necessitating a sufficient infrastructure for deploying and evaluating them.

5.4 Blogs

Weblogs or 'blogs' are an internet phenomenon originating in the late 1990s: Websites that continually publish new articles on their front page, written by one individual or a group of people. Blog entries can be tied to their author; linking between entries is supported in the form of comments or so-called 'trackback links', in which the author of another blog refers in his entry to the original entry.

Blogs can take many forms: personal blogs usually form a sort of diary of the owner's thoughts and interests. Topical blogs are usually edited by several people and publish information about a specific topic. Corporate blogs may give the executives and other employees a platform for publishing news articles. A number of service provides exist on the internet that allow one to create a blog free of charge; examples are BlogSpot[19] (now owned by Google) and LiveJournal[20].

Another typical feature is the so-called blogroll: A list of other blogs the author reads regularly. This may be used to determine social links between authors, but it is not universally adopted.

Obtaining the archives of a blog is difficult in general, as no standardized format or repository exists. It may require collaboration with the individual authors.

[17] http://c2.com/cgi/wiki?WikiWikiWeb, last visit on 2005/03/08.
[18] http://www.wikipedia.org/, last visit on 2005/08/03.
[19] http://www.blogspot.com/, last visit on 2005/03/09.
[20] http://www.livejournal.com/, last visit on 2005/03/09.

5.5 Messenging systems

There are a number of messenging systems that are sufficiently similar to each other to be grouped under one heading; examples are email mailing lists, Usenet, and web forums. These systems are among the oldest collaborative electronic mediums; however, as articles are often written 'off the cuff' and cannot be revised, they are often lacking in quality. Mailing list archives can be a valuable repository of knowledge, but separating the wheat from the chaff is notoriously difficult.

Messenging systems are usually characterized by a tree structure of links between individual documents. A further speciality is that the polarity of the link structure is reversed: A follow-up article is usually not a sign of support, but a sign of disagreement or a sign that the original article is lacking information.

Mailing list archives for public mailing lists are usually rather easy to obtain; they are either published on a web site, or can be constructed trivially by subscribing to the mailing list in question for a few months; Usenet archives are obtainable in a similar manner. Obtaining the archives of a web forum may require collaboration with the forum's owner.

6 Additional concerns: Privacy, Anonymity and Plausible Deniability

Privacy matters quickly become a concern as more and more information can be tied to a specific person, especially when this information concerns said person's interests and social ties to other people. Therefore, a user of such a system needs to be aware of the information that is available on him and how it is used. He needs to know that certain acts of his constitute a publication that all others can see.

We propose techniques that actively make use of that fact that information can be tied to a specific user and can be made available to others – in order to identify relevant users and their content. As such, we believe that they should only be used in environments where the information is publicly available anyway. We oppose to them being used in combination with information gathering techniques such as evaluating browsing histories: In such an environment, the user has no direct control about the information that is published about him.

In other applications, measures need to be taken to ensure anonymity of the individual users and, preferably, plausible deniability (an individual user can plausibly deny that a specific piece of information originated from him.) How to employ our techniques in such an environment is not subject of this thesis.

7 Evaluation

Unfortunately, social information retrieval is not yet a common problem, and there are no suitable standard corpora available. We intend to perform known-item retrieval on two out of three locally available corpora in order to compare the performance of social information retrieval to classic methods. The following corpora are at our disposal for the evaluation of our techniques:

- A collection of 25 years of conference proceedings of the annual ACM SIGIR (Special Interest Group on Information Retrieval) conference

  This corpus contains author and title information about every paper published in the proceedings of the SIGIR conference, as well as a full-text index. References pointing to other papers in the corpus are also present, as well as the context (surrounding text) of said references.

  Social ties can be inferred from coauthorship information; links between individual papers can either be in the form of references, or textual similarity.

- The German Wikipedia

  Complete database dumps of the German Wikipedia, complete with all revisions, are available for download. At 12GB (compressed), the size of the German database is still manageable, in contrast to the English version. The availability of database dumps and machine-readable markup of individual entries ensures easy access to the relevant information, without the need for elaborate parsing and information extraction from natural-language text.

  The strength of social ties between two users can be assessed in two ways: Firstly, by determining which users revised the same articles (amounting to coauthorship), and secondly by detecting which user revised another's talk page (amounting to direct communication between two users.) Links between articles can be qualified by textual similarity or by inter-article links (which can easily be parsed from Wiki markup.)

- A mailing list archive

  For further evaluation, a complete mailing list archive of the 'origami-l' mailing list[21] from April 1997 to the present date is also available; the archive was collected by the author.

The applications identified in section 4 are expressed by querying the associative network in different ways:

[21] http://origami.kvi.nl/, last visit on 2005/05/02.

By allowing simple keyword queries, we provide the user with an interface similar to conventional information retrieval systems; however, we expect our system to produce an improved ranking by taking social information into account (Section 4.1.)

If the user himself is part of the associative network, personalized search results are produced by combining a keyword query and a query for the user's own node in the network. This way, results from the user's immediate vicinity are ranked higher than results from other parts of the network. (Section 4.2.)

By querying for his own node, the user can determine a subset of nodes that form his immediate vicinity. This subset can be used to limit the scope of a search, for example in distributed systems. (Section 4.4.)

Since no standard test collections exist for the task at hand, we intend to use known-item retrieval for evaluation. The basic setting for known-item retrieval is a user that wants to find a document he has seen before. Ideally, a system would only produce one relevant document in such a setting – namely the document the user was looking for in the first place. Performance will be measured using precision (percentage of results retrieved that are relevant to the query) and recall (percentage of relevant results retrieved.)

For the SIGIR corpus, the context of references will aid us in formulating queries; as this information is not used in retrieval, it provides an independent way of evaluating the search effectiveness.

8 Structure of the diploma thesis

The diploma thesis will comprise the following chapters:

1. Introduction
2. The social information retrieval concept
    a) Definition
    b) Requirements analysis
    c) Related work
3. Techniques for social information retrieval
    a) Spreading Activation Search
    b) Flow-based communities
    c) Vector-space retrieval
4. Evaluation: Setting, experiments and results
5. Conclusion

9 Schedule

We estimate the following timeframe for the individual sub-tasks:

- Developing the social information retrieval concept, requirements analysis (4 weeks)
- Preparing the evaluation datasets (3 weeks)
- Implementing the social information retrieval concept (7 weeks)
- Experiments and evaluation (4 weeks)
- Further research into related systems, conclusion (4 weeks)
- Revising the thesis (2 weeks)

References

John R. Anderson. *The Architecture of Cognition*. Cognitive Science Series. Harvard University Press, 1983. ISBN 0-674-04425-8.

F. Anklesaria, M. McCahill, P. Lindner, D. Johnson, D. Torrey, and B. Albert. The Internet Gopher Protocol (a distributed document search and retrieval protocol). RFC 1436 (Informational), March 1993. URL http://www.ietf.org/rfc/rfc1436.txt.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

Krishna Bharat. Searchpad: explicit capture of search context to support web search. *Comput. Networks*, 33(1-6):493–501, 2000. ISSN 1389-1286.

Maciej Ceglowski, Aaron Coburn, and John Cuadrado. Semantic search of unstructured data using contextual network graphs, 2003. URL http://research.nitle.org/papers/Contextual_Network_Graphs.pdf.

F. Crestani. Application of spreading activation techniques in information retrieval. *Artif. Intell. Rev.*, 11(6):453–482, 1997. ISSN 0269-2821.

Raymond D'Amore. Expertise community detection. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 498–499, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-881-4.

Gary William Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160. ACM Press, 2000. ISBN 1-58113-233-6.

Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-organization and identification of web communities. *Computer*, 35(3):66–71, 2002. ISSN 0018-9162.

Gary William Flake, Kostas Tsioutsiouliklis, and Leonid Zhukov. Methods for mining web communities: Bibliometric, spectral, and flow. In Alexandra Poulovassilis and Mark Levene, editors, *Web Dynamics*, chapter 4, pages 45–68. Springer Verlag, 2004. ISBN 3-540-40676-X. URL http://research.yahoo.com/publications/4.pdf.

William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval. Data Structures & Algorithms*. Prentice Hall, 1992.

Jill Freyne and Barry Smyth. An experiment in social search. In Wolfgang Nejdl and Paul De Bra, editors, *Adaptive Hypermedia and Adaptive Web-Based Systems, Third International Conference, AH 2004, Eindhoven, The Netherlands, August 23-26, 2004, Proceedings*, volume 3137 of *Lecture Notes in Computer Science*, pages 95–103. Springer, 2004. ISBN 3-540-22895-0.

Melanie Gnasa, Sascha Alda, Jasmin Grigull, and Armin B. Cremers. Towards virtual knowledge communities in peer-to-peer networks. In Jamie Callan, Fabio Crestani, and Mark Sanderson, editors, *Distributed Multimedia Information Retrieval*, volume 2924 of *Lecture Notes in Computer Science*, pages 143–155. Springer, 2003. URL http://www.springerlink.com/index/NUR92TH9821N5TPJ.

Melanie Gnasa, Markus Won, and Armin B. Cremers. Three pillars for congenial web search. continuous evaluation for enhancing web search effectiveness. *Journal of Web Engineering*, pages 252–280, 2004. ISSN 1540-9589. URL http://www.informatik.uni-bonn.de/~won/Download/wwwjournal2004.pdf.

R. Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 403–412, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-844-X.

Henry Kautz, Bart Selman, and Mehul Shah. The hidden web. *AI Magazine*, 18(2): 27–36, 1997a. URL http://www.cs.washington.edu/homes/kautz/referralweb/doc/aimag.pdf.

Henry Kautz, Bart Selman, and Mehul Shah. Referral web: combining social networks and collaborative filtering. *Commun. ACM*, 40(3):63–65, 1997b. ISSN 0001-0782.

Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. GroupLens: applying collaborative filtering to Usenet news. *Commun. ACM*, 40(3):77–87, 1997. ISSN 0001-0782.

F. W. Lancaster. *Information Retrieval Systems: Characteristics, Testing, and Evaluation*. Wiley, New York, 1968.

Larry Masinter and Erik Ostrom. Collaborative information retrieval: Gopher from MOO. In *Proceedings of INET '93*, 1993. URL http://larry.masinter.net/MOOGopher.pdf.

David Mattox, Mark T. Maybury, and Daryl Morey. Enterprise expert and knowledge discovery. In *Proceedings of the HCI International '99 (the 8th International Conference on Human-Computer Interaction)*, pages 303–307, Mahwah, NJ, USA, 1999. Lawrence Erlbaum Associates, Inc. ISBN 0-8058-3392-7. URL http://www.mitre.org/work/tech_papers/tech_papers_00/maybury_enterprise/maybury_enterprise.pdf.

Mark Maybury, Ray D'Amore, and David House. Expert finding for collaborative virtual environments. *Commun. ACM*, 44(12):55–56, 2001. ISSN 0001-0782.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, November 1999. URL http://dbpubs.stanford.edu:8090/pub/1999-66.

Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: extracting usable structures from the web. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 118–125, New York, NY, USA, 1996. ACM Press. ISBN 0-89791-777-4. URL http://www.pitkow.com/docs/1996-CHI-Silk.pdf.

Nicholas C. Romano, Jr, Dmitri Roussinov, Jay F. Nunamaker, Jr, and Hsinchun Chen. Collaborative information retrieval environment: Integration of information retrieval with group support systems. In *HICSS '99: Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 1*, page 1053, Washington, DC, USA, 1999. IEEE Computer Society. ISBN 0-7695-0001-3.

Gerard Salton and Chris Buckley. On the use of spreading activation methods in automatic information retrieval. In *Proceedings of the ACM SIGIR*, Grenoble, France, June 1988. URL http://doi.acm.org/10.1145/62437.62447.

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999. ISSN 0163-5840.

Alan F. Smeaton, Gary Keogh, Cathal Gurrin, Kieran McDonald, and Tom Sødring. Analysis of papers from twenty-five years of sigir conferences: what have we been doing for the last quarter of a century? *SIGIR Forum*, 36(2):39–43, 2002. ISSN 0163-5840. URL http://portal.acm.org/citation.cfm?id=792556.

Etienne Wenger. How we learn. Communities of practice. The social fabric of a learning organization. *Healthcare Forum Journal*, 39(4):20–6, 1996. URL http://www.ewenger.com/pub/pubhealthcareforum.htm.

T. D. Wilson. On user studies and information needs. *Journal of Librarianship*, 37(1): 3–15, 1981. URL http://informationr.net/tdw/publ/papers/1981infoneeds.html.

T. D. Wilson. Information needs and uses: fifty years of progress. In B. C. Vickery, editor, *Fifty years of information progress: a Journal of Documentation review*, pages 15–51. Aslib, London, 1994. URL http://informationr.net/tdw/publ/papers/1994FiftyYears.html.

Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, New York, NY, USA, 1996. ACM Press. ISBN 0-89791-792-8.

Atsushi Yamamoto, Daisuke Asahara, Tomoko Itao, Satoshi Tanaka, and Tatsuya Suda. Distributed pagerank: A distributed reputation model for open peer-to-peer networks. In *SAINT-W '04: Proceedings of the 2004 Symposium on Applications and the Internet-Workshops (SAINT 2004 Workshops)*, page 389, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2050-2.

Dawit Yimam-Seid and Alfred Kobsa. Expert finding systems for organizations: Problem and domain analysis and the DEMOIR approach. *Journal of Organizational Computing and Electronic Commerce*, 13(1):1–24, 2003.

Bin Yu and Munindar P. Singh. Searching social networks. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 65–72, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-683-8. URL http://portal.acm.org/citation.cfm?id=860587.