Peer-to-Peer Information Search: Semantic, Social, or Spiritual?*

Matthias Bender, Tom Crecelius, Mouna Kacimi,
Sebastian Michel, Josiane Xavier Parreira, Gerhard Weikum

Max-Planck Institute for Informatics
Saarbruecken, Germany
{mbender, tcrecel, mkacimi, smichel, jparreir, weikum}@mpi-inf.mpg.de

Abstract

We consider the network structure and query processing capabilities of social communities like bookmarks and photo sharing communities such as del.icio.us or flickr. A common feature of all these networks is that the content is generated by the users and that users create social links with other users. The evolving network naturally resembles a peer-to-peer system, where the peers correspond to users. We consider the problem of query routing in such a peer-to-peer setting where peers are collaborating to form a distributed search engine. We have identified three query routing paradigms: semantic routing based on query-to-content similarities, social routing based on friendship links within the community, and spiritual routing based on user-to-user similarities such as shared interests or similar behavior. We discuss how these techniques can be integrated into an existing peer-to-peer search engine and present a performance study on search-result quality using real-world data obtained from the social bookmark community del.icio.us.

1 Introduction

Peer-to-peer (P2P) information management and search is intriguing for scalability and availability. In addition, a P2P network would be a natural habitat for exploiting the "social wisdom" of its users. We envision a P2P system where each user runs a peer computer (e.g., on her PC, notebook, or even cell phone) and shares information within a large community. Each peer would be a full-fledged data management system for the user's personal information, scholarly work, or data that the user may harvest (and cache) from Internet sources such as news, blogs, or specialized Web portals. Each peer would also have a local search engine, which could be very powerful (e.g., using advanced NLP, machine learning, and ontologies), given that it operates on the user's relatively small-sized information collection on a dedicated computer, and could be highly customized to the user's individual interests and behavior. The Minerva platform developed in our group [4] follows this paradigm; other projects along the same lines include, for example, pSearch [37], Alvis [23], and BestPeers [18].

As a futuristic application scenario consider millions of users who use their mobile devices to record photos and videos of all kinds of real-world events ranging from business meetings to vacation trips. Such digital-perception information can be easily annotated with speech and device-generated metadata such as GPS and time

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*This work is partially supported by the EU project SAPIR.

coordinates. Moreover, all this data could be made accessible on a P2P network instantaneously, for search and further annotation - so-called "social tagging" - by other users. For example, thousands of tourists on the Forum Romanum can immediately share their photos and annotations, so that an uninitiated tourist could immediately receive explanations about some lesser known remains from the annotations of other, more knowledgeable users. The P2P search and social-networking technology that underlies such a scenario would be embedded in the application software and be virtually invisible to the end-users.

In such P2P settings, queries would first be executed locally, on the peer where the query is issued. This would utilize the locally available information and powerful, personalized search capabilities. In some settings, this may be a local cache of annotated photos or MP3 files; in others, it could be a collection of personally relevant Web pages that have been compiled by thematically focused crawling and subscriptions to feeds. If the local search does not return satisfactory results, the peer should consider forwarding the query to a small number of judiciously chosen other peers. This step towards collaborative search is known as the *query routing* decision. It should consider both the expected benefits of obtaining better information from other peers and the communication and execution costs of involving these peers. The literature on P2P information retrieval and other forms of distributed IR contains many proposals for query routing strategies; see, e.g., [25, 14, 16, 22, 30, 5, 26, 3].

The routing decision is usually driven by various forms of precomputed (and incrementally maintained) routing indices, peer-content synopses, or distributed directories, which in turn can influence the topology of the P2P overlay network leading to so-called semantic overlay networks (SONs) [11, 31, 2, 21, 12, 1].

In the current paper, we do not make any assumptions about this infrastructure or the overlay topology, and rather assume that the query routing decision has all the information about other peers that it needs and chooses peers solely by benefit/cost considerations. We will disregard the cost aspects for this paper and focus on the much less explored benefit issues.

We investigate three broad families of strategies:

- *Semantic query routing*: The peers to which a query is forwarded are chosen based on the *content similarity* between the query and the data held by the candidate target peers (or the corresponding peer synopses).
- *Social query routing*: The target peers are chosen based on *social relationships* like the explicitly listed friends of the query initiator or peers that belong to the same explicit groups.
- *Spiritual query routing*: The target peers are chosen based on *behavioral affinity* such as high overlap in tag usage, bookmarked pages, or commenting and rating activity. This aims to capture "brothers in spirit", hence the name.

We refer to the first family as "semantic" as the content comparison could take into account metadata (e.g., schema mappings), ontology-based similarities, and other aspects that go beyond purely syntactic or statistical measures. For simplicity, the current paper considers only keyword queries (referring to text terms or user-provided tags) and consequently uses simple measures of (IR-style) statistical similarity, but the approach could be enriched and generalized. The second and the third approach are closely related and could be easily confused. We refer to "social search" when explicit friendship or other social-networking relations are used, and we refer to "spiritual search" when considering users' tagging, bookmarking, rating, and other behaviors.

This paper discusses how these three approaches can be used in P2P query routing, and how effective they are for delivering high-quality results. As we consider keyword queries, we will use IR quality measures like precision and recall. We also present hybrid strategies that combine elements from both semantic and social or semantic and spiritual search. The rest of the paper is organized as follows. Section 2 briefly reviews the state of the art on P2P information search and its relation to social networks. Section 3 presents the Minerva system architecture, which is our testbed and serves as a representative of the general architectures to which our work applies. Section 4 introduces our query routing strategies in more detail. Section 5 presents an experimental

comparison of different strategies, using data extracted from the popular social-tagging site *del.icio.us*. Section 6 points out lessons learned and future work.

2 Related Work

One of the fundamental functionalities that a P2P information system must provide is to identify the most "appropriate" peers for a particular query, i.e., those peers that are expected to locally hold high-quality results for the query. This task is commonly referred to as query routing, sometimes also as resource or collection selection. We stress that query routing is more challenging than it may appear at first sight: the set of peers to be contacted is not simply the set of all peers that store relevant index data. Such a set could contain a very large number of peers and contacting all of them would be prohibitive. While there exist a number of approaches for query routing in the literature on distributed IR — e.g., CORI [9], GlOSS [16], and methods based on statistical language models [34] — these were typically designed for a stable and rather small set of collections (e.g., in the context of metasearch engines). These techniques usually assume that the document collections are disjoint, which is a rather unrealistic assumption in P2P systems where the peers are compiling their content (e.g., by crawling the Web) at their discretion. In [5, 27] we have proposed the usage of overlap aware query routing strategies. The proposed methods use compact data synopses such as Bloom filters or hash sketches to estimate the mutual overlap between peers to avoid querying peers that provide basically the same information, which would waste both processing power and network resources.

The statistical summaries describing a peer are usually organized on a per-term basis, indicating the expected result quality of a peer's collection for a given term. This limitation is considered unavoidable, as statistics on all term pairs would incur a quadratic explosion, leading to a breach with the goal of scalability. On the other hand, completely disregarding correlations among terms is a major impediment: for example, consider the following extreme scenario. Assume peer p_1 contains a large number of data items for each of the two terms a and b separately, but none that contains both a and b together. Judging only by per-term statistics, state-of-the-art query routing approaches would reach the conclusion that p_1 is a good candidate peer for the query $\{a, b\}$, whereas the actual result set would be empty. In [26, 6], we present a routing method that uses multi-key statistics to improve the query routing performance. We propose the usage of a distributed query-log analysis to discover frequently co-occurring keys (terms) that are candidates for being considered as additional keys in the distributed directory. To decrease the directory load, we introduce a pruning technique to avoid considering unnecessary key-sets.

Social networks have recently emerged in P2P systems to address several issues such as improving content discovery [13, 8, 19], reducing latency and speeding up downloads [32, 38, 35], and designing trust models [24, 17]. In the following, we briefly present some approaches towards P2P search.

Pouwelse et al. [32] propose Tribler, a social-based P2P overlay on top of BitTorrent. It connects peers based on their similar "tastes" instead of considering similar files. Thus, peers exploit their social links and invoke the help of their friends to improve content discovery and download cost. Similarly, Fast et al. [13] propose using user interests to build social groups in a P2P network. Users sharing the same type of files are connected to each other even though their contents do not overlap. The main goal of this approach is to capture important aspects of download behavior by connecting peers to the potential providers of their required files.

Other social P2P networks are based on peer request traces. A peer uses request relationships to other peers to construct social links to them. Sripanidkulchai et al. [35] implement a performance enhancement layer on top of the flooding-based content location mechanism of Gnutella. Each peer creates and maintains its shortcuts list based on its request trace. Shortcuts are ranked according to some metrics such as the probability of providing relevant content, latency of the path to the shortcut, available path bandwidth, shortcut load, etc. The work presented by Tempich et al. [38] considers query traces to create a human social network. It defines a query routing strategy in which peers observe which queries are successfully answered by other peers and remember

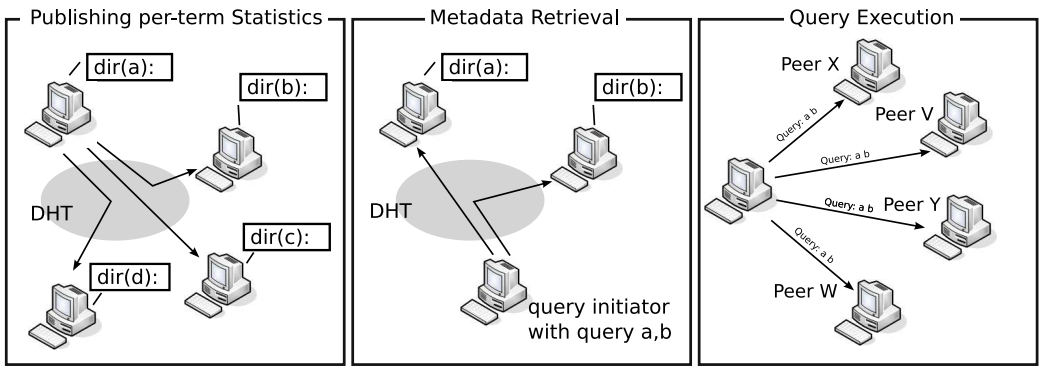

Figure 1: Metadata Dissemination, Query Routing, and Query Execution in Minerva.

these peers in future query routing decisions.

Borch et al. [8] present a social P2P search infrastructure which groups peers based on the similarity of their keyword searches. The authors describe different application scenarios including distributed bookmark sharing in which users create bookmarks, describe them using tags, and share them with friends and colleagues. The basic idea is to send queries to peers likely to have interesting resources. Khambatti et al. [19] introduce the notion of peer communities that consist of active peers involved in sharing, communicating, and promoting common interests. These communities are self-organizing using distributed formation and discovery algorithms.

3 Minerva

We have developed a P2P Web search engine coined Minerva, released as open source and available under *http://www.minerva-project.org*. We envision a network of peers, each with a local index and a local query processor, that are crawling the Web independently, for example, to harvest blogs or scientific publications according to the user's thematic profile. Minerva maintains a metadata directory that is layered on top of a distributed hash table (DHT) [36, 33]. It holds very compact, aggregated summaries of the peers' local indexes and only to the extent that the individual peers are willing to disclose. A query initiator selects a few most promising peers based on their published per-term summaries, e.g., by executing a distributed top-k algorithm like [10, 28]. Subsequently, it forwards the complete query to the selected peers which execute the query locally. This query execution does not involve a distributed top-k query execution since each peer maintains a full-fledged local index with all information necessary to execute the query locally. Finally, the results from the various peers are combined at the querying peer into a single result list.

Figure 1 illustrates the Minerva approach. First, every peer publishes per-term summaries (*Posts*) of its local index to the directory. The DHT (and its replication mechanism) determines the peer(s) currently responsible for this term. This peer (or these peers in the case of replication) maintains a *PeerList* of all postings for this term from across the network. Posts contain contact information about the peer who posted a summary together with statistics to calculate IR-style measures for a term (e.g., the size of the inverted list for the term, the average score for the term's inverted list entries, or other statistical measures). These statistics are used to support the query routing decision, i.e., determining the most promising peers for a query.

Minerva facilitates easy integration of new query routing strategies, like the ones proposed in this paper. For instance, users' bookmarks can be crawled and indexed, and their terms can then be posted to the distributed metadata directory. Similarly, tags used to describe the bookmarks can be stored in the directory. This supports semantic query routing. For the social and spiritual query routing, the Minerva framework can be extended by keeping, at each peer, a list of peers that are related either by social relationship or behavioral affinity. Note that

these lists tend to be very small, relative to the size of the network; so the approach scales up well.

In the spirit of social tagging communities, users can manually add arbitrary attribute-value annotations by a single mouse click. For example, users might rate Web pages or blogs with annotations such as *rating=5*. Additional annotations may be automatically generated from the content, such as *author=weikum* or *conference=ICDE*. These annotations are also indexed and become part of the directory; so users can explicitly query for documents with *rating=5* and also combine such conditions with query keywords.

4 Query Routing in Social P2P Networks

4.1 Semantic Query Routing

The peers to which a query is forwarded are chosen based on the *content similarity* between the query and the data held by the candidate target peers (or the corresponding peer synopses). The query is represented by its keywords – *terms* in IR jargon –; the data of a peer can be represented by its terms or its tags or a combination of both. With each term and each tag we can also associate some precomputed frequency statistics, e.g., how often a term or tag has been used by a given peer and how often it is used in the overall P2P network. Following query-routing terminology, we refer to the total frequency of tag or term t at peer p_j as the *document frequency* $df_j(t)$; this is the number of bookmarked pages in p_j's collection that contain or are tagged with term/tag t.

Semantic query routing estimates the benefit for different candidates based on the sum of document frequencies for the query terms (as determined by the best entries from the term- or tag-specific directory entries fetched via DHT lookups), and chooses the highest-ranked peers according to this measure. Alternatively, one could also employ more sophisticated methods such as CORI [9] that uses, in addition to the document frequency, several dampening and smoothing techniques partially based on the notion of collection frequencies, i.e., the number of peers that have bookmarked pages that contains a particular term.

4.2 Social Query Routing

The target peers are chosen based on *social relationships*. We assume that there is an explicit *friends* relation among peers, and we choose target peers for forwarding a query issued at peer p_j to be the "best" friends of p_j, provided the degree of friendships are quantified (e.g., based on the frequency of interactions between peers in the recent past). If there is no quantitative measure for friendship strength, then we simply choose a random subset of friends when we want to limit the number of target peers, or all friends when there is no limit.

4.3 Spiritual Query Routing

The target peers are chosen based on *behavioral affinity* such as high overlap in tag usage, bookmarked pages [7], or commenting and rating activity. We could use an information-theoretic measure, the Kullback-Leibler divergence (relative entropy) [20], on the tag frequency distributions of a peer's bookmarked pages (possibly combined with rating information), and would quantify the similarity for each pair of peers. We can then use such a similarity measure to cluster peers that are spiritually close to each other. A simpler approach with the same intention considers the overlap in the bookmarked pages among peers. This can be efficiently computed in a P2P environment using distributed algorithms on compact synopses like Bloom filters [5, 27]. Spiritual query routing for a query initiated at peer p_i then chooses the peers p_j with the highest estimated $overlap(p_i, p_j)$.

4.4 Hybrid Strategies

All the aforementioned routing strategies can be combined into hybrid methods. Here we outline only some straightforward approaches and leave more sophisticated combinations for future work. The goal of peer selection is to identify the top-k peers for a particular query. A hybrid approach would select k_i peers with strategy

S_i so that $\sum_i k_i = k$. The choice of the single k_i values is a nontrivial problem (cf. [29]). A simple approach would, for example, use $k_1 = k_2, ...$ and a round-robin selection.

Combining the social routing strategy with a spiritual routing strategy would, for instance, decrease the risk of obtaining mediocre results when the query does not fit with the friends' thematic interests in a purely social routing strategy.

4.5 Orthogonal Issues

Besides the aforementioned query routing concepts that aim to find promising peers for a particular information need, an overlap-aware technique [5, 27] can be employed to eliminate redundancy in the query evaluation. For instance, it does not make sense to query both peers A and B if it is known that both have (almost) the same information or A's collection is a subset of B's collection.

5 Experiments

5.1 Data Collection

We have crawled parts of del.icio.us[1] with a total of $13,515$ users, $4,582,773$ bookmarks, and $152,306$ friendship connections. In addition, we have actually crawled and indexed the actual HTML pages where the bookmarks point to, giving us the possibility to execute both term-based and tag-based queries.

Each peer in our experiments corresponds to exactly one user. The local collection of a peer consists of the bookmarked pages, including their actual contents, and the user-provided tags for each page.

5.2 Queries

For the workload we needed realistic queries and their association with specific users. Query logs with this kind of information are not publicly available. Therefore, we generated queries based on the users' tags in a way that the queries reflect the user interests. For a particular user we consider those tags that frequently co-occur for the same bookmarks.

More precisely, to generate the benchmark queries, we first identified the top ξ users in terms of bookmark-set cardinalities. Then we considered those tag pairs that were used together at least ζ times and not more than ψ times by the selected users. The first constraint is needed to eliminate rare tag pairs. The second constraint is used to eliminate tag pairs that have a stopword character. For our experiments we chose $\xi = 5$, $\zeta = 200$, and $\psi = 900$. Using this technique we identified 24 queries with two tags, such as "music media", "web design", "mac apple", and "tech reference".

5.3 Quality Measures

There are no standard queries and no relevance assessments available for the pages bookmarked in del.icio.us. We consider two different approaches for defining some notion of "ground truth": a hypothesized ideal search result to which our strategies can be compared.

(i) As a first approach, we use pages bookmarked by the query initiator as ground truth. Consider a multi-keyword query $Q = q_1, q_2, ..., q_m$. The query initiator retrieves the top-k pages from each of the peers selected during the query routing phase. Then, to estimate the quality of the retrieved pages, the initiator compares the obtained results with the pages she has bookmarked and tagged with tags $q_1, q_2, ..., q_m$. The rationale behind this evaluation is that the fact that a user has bookmarked a page can be interpreted as relevance judgment.

[1] *http://del.icio.us*

(ii) As an alternative approach, which is independent of the query initiator, we consider all pages that are bookmarked in the system and tagged (by some user) with all the query keywords as relevant. The goal for the query execution then is to maximize the number of results from this pool of relevant pages.

For the first approach, the "relevance judgments" highly depend on the query initiator. Thus, we have to select as query initiators "power users" with a sufficiently large number of bookmarks. We first select a query by choosing a frequent tag pair. Then we rank peers that have at least 50 friends according to the number of bookmarks that are tagged with the chosen pair. For each query (i.e., keyword pair) we consider the top-5 peers as query initiators, i.e., we execute the same query five times to remove the influence of an accidentally bad choice for one of the initiators.

The second approach allows for relevance assessment that is independent of the query initiator, whereas the first approach depends on the choice of the query initiator. However, the social and the spiritual routing strategies depend on the query initiator anyway, as, for instance, executing a query related to pop music on a peer that is primarily interested in soccer would not return good results by design.

Once a peer receives an incoming query request, it executes the query locally and returns *all* bookmarked pages that are *tagged* with the keywords in the query. In a real-world system one would try to return only the top-k results by some meaningful ranking. However, as we deal with personalized search here, it is not straightforward to apply a standard scoring model. Therefore, we let peers return all bookmarked pages that match the query.

The same situation occurs when we merge the result lists returned by the queried peers: as there is no widely agreed merging strategy, we assess the quality of the union of the returned results.

5.4 Strategies under Comparison

For multi-keyword queries of the form $Q = \{t_1, .., t_m\}$ we evaluate the retrieval quality, measured by recall (relative to the ground truth explained in the previous subsection), of the following strategies:

- **Semantic Routing based on Tags:** We rank peers according to the sum of document frequencies, i.e., the score of a peer p_i is given by $\sum_{t \in Q} df_i(t)$ where $df_i(t)$ is the number of bookmarks in peer p_i's collection that are tagged with t, cf. Section 4.1.

- **Semantic Routing based on Terms:** We rank peers according to the sum of document frequencies, similar to the tag based semantic routing, but here we consider terms instead of tags.

- **Social Routing:** We let the query initiator send the query to the top friends where the friends are ranked according to the number of bookmarks they have.

- **Spiritual Routing:** For spiritual closeness we consider the overlap in the bookmarks.

- **Hybrid between Semantic and Spiritual Routing:** This hybrid strategy combines the routing results (peer rankings) obtained from the semantic and spiritual routing strategies in a round-robin manner, ignoring duplicates.

- **Hybrid between Semantic and Social Routing:** This is a combination of the semantic and the social routing results using a round-robin selection process, ignoring duplicates.

- **Hybrid between Spiritual and Social Routing:** This is a combination of the spiritual and the social routing results using a round-robin selection process, ignoring duplicates.

5.5 Experimental Results

Figure 2 shows the average recall for the benchmark with 120 queries (24 distinct queries, each issued by 5 different peers) when considering the query initiator's bookmarks as the ground truth. The semantic routing strategy is the clear winner. The spiritual routing strategy performs reasonably well but cannot reach the performance of the semantic routing strategy. For instance, when asking 10 peers, the semantic routing strategy achieves a recall of nearly 16% whereas the spiritual strategy achieves approximately 8% recall. The social routing strategy performs worse than all other strategies. Surprisingly, the term-based semantic routing strategy performs poorly. This is probably due to the particular nature of the queries that have been created based on the most popular tags as many tags are not "appropriate" search terms. Examples are "Task Organizing" tags [15] like "toread" or "jobsearch". [15] gives a nice overview on the different functions that tags can have.

The relative order of the hybrid strategies follows that of the pure strategies: the semantic-spiritual strategy is the best hybrid strategy, followed by the semantic-social strategy, and the spiritual-social strategy performs worst but still better than the purely social strategy.

Figure 3 shows similar results for the second choice of ground truth with bookmarked pages that are tagged with the query words as relevant. The results confirmed our findings from the first experiment; so no further discussion is needed here.

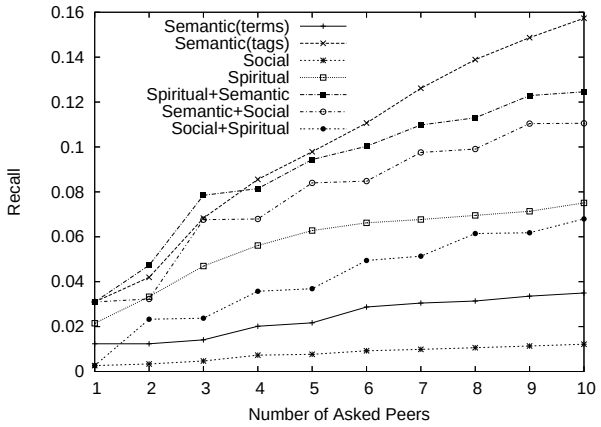

Figure 2: Average Recall: considering the query initiator's bookmarks that match the query tags as relevant.

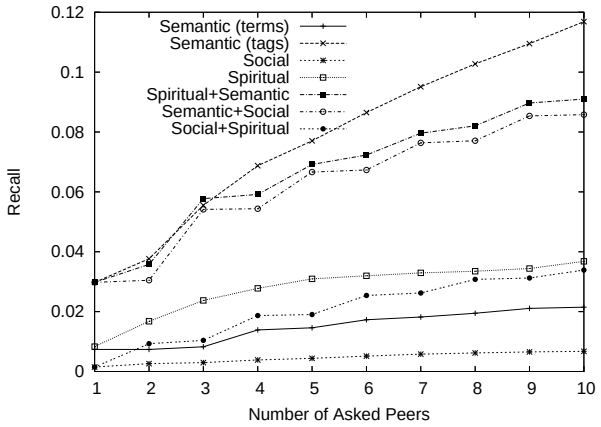

Figure 3: Average Recall: considering all bookmarks that match the query tags as relevant.

6 Lessons Learned and Future Work

Our experiments have shown that the semantic routing strategies that use per-tag peer summaries are superior to all other strategies. The social routing strategy performed very poorly in our experiments, and it is disappointing to see that it provided hardly any relevant results.

To understand this poor performance we have analyzed the content overlap among peers that are related by friendship connections. For each user, we have calculated the overlap between her bookmarks and the bookmarks from her friends. It turned out that the overlap is surprisingly small: considering only users that have at least one friend, the mean value is about 7%, i.e., half of the peers share less than 7% of their bookmarks with their neighbors. The minimum overlap observed was 0.03 %, the first and third quartiles were 2.8% and 14.5%, respectively. Theses low numbers partially explain the bad performance of the social routing strategy. In our experiments, for half of the users, a recall of at most 7% would be obtained if we had asked all their friends. Since we limited the number of friends queried to 10, the obtained recall was even lower.

We believe that this phenomenon is due to the particular usage of the friend relationships in del.icio.us. It seems that users establish a new friendship connection when the bookmarks tagged by the new friend are considered as interesting, and then the user does not care anymore about tagging the same pages. This interesting feature of such networks may need further exploration.

Note that the social routing strategy does not require any global information like the semantic strategy and the spiritual strategy. The semantic strategy needs a global mapping from tags (or terms) to per-peer summaries that cause some maintenance cost (to update the DHT-based directory). The spiritual routing strategy requires continuous peer meetings to learn about thematically close peers, although these information exchanges could probably be piggybacked on messages that are sent anyway on behalf of user queries.

Our intention in this paper was to outline our framework for semantic, social, and spiritual query routing, identify technical issues, and shed some light into the experimental behavior of these P2P routing strategies within social networks. Our findings clearly dampen the optimism about social networks being able to boost search result quality in a P2P network. More traditional content-oriented strategies were found to be way superior. However, our observations and insights are clearly preliminary at this point, and should stimulate further research in this area.

References

[1] K. Aberer and P. Cudré-Mauroux. Semantic overlay networks. In *VLDB*, page 1367, 2005.

[2] K. Aberer, P. Cudré-Mauroux, M. Hauswirth, and T. V. Pelt. Gridvine: Building internet-scale semantic overlay networks. In *International Semantic Web Conference*, pages 107–121, 2004.

[3] R. Baeza-Yates, D. Puppin, and R. Perego. Incremental caching for collection selection architectures. In *Infoscale*, 2007.

[4] M. Bender, S. Michel, J. X. Parreira, and T. Crecelius. P2p web search: Make it light, make it fly (demo). In *CIDR*, pages 164–168, 2007.

[5] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Improving collection selection with overlap awareness in p2p search engines. In *SIGIR*, pages 67–74, 2005.

[6] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. P2P content search: Give the web back to the people. In *5th International Workshop on Peer-to-Peer Systems (IPTPS 2006)*, 2006.

[7] M. Bender, S. Michel, G. Weikum, and C. Zimmer. Bookmark-driven query routing in peer-to-peer web search. In *Workshop on Peer-to-Peer Information Retrieval*, 2004.

[8] N. Borch. Social peer-to-peer for social people. In *The International Conference on Internet Technologies and Applications*, 2005.

[9] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *SIGIR*, pages 21–28, 1995.

[10] P. Cao and Z. Wang. Efficient top-k query calculation in distributed networks. In *PODC*, pages 206–215, 2004.

[11] A. Crespo and H. Garcia-Molina. Semantic overlay networks for p2p systems. In *AP2PC*, pages 1–13, 2004.

[12] C. Doulkeridis, K. Nørvåg, and M. Vazirgiannis. The sowes approach to p2p web search using semantic overlays. In *WWW*, pages 1027–1028, 2006.

[13] A. Fast, D. Jensen, and B. N. Levine. Creating social networks to improve peer-to-peer networking. In *KDD*, pages 568–573, 2005.

[14] N. Fuhr. A decision-theoretic approach to database selection in networked ir. *ACM Trans. Inf. Syst.*, 17(3):229–249, 1999.

[15] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[16] L. Gravano, H. Garcia-Molina, and A. Tomasic. Gloss: Text-source discovery over the internet. *ACM Trans. Database Syst.*, 24(2):229–264, 1999.

[17] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *ESWC*, pages 411–426, 2006.

[18] P. Kalnis, W. S. Ng, B. C. Ooi, and K.-L. Tan. Answering similarity queries in peer-to-peer networks. *Inf. Syst.*, 31(1):57–72, 2006.

[19] M. Khambatti, K. D. Ryu, and P. Dasgupta. Structuring peer-to-peer networks using interest-based communities. In *DBISP2P*, pages 48–63, 2003.
[20] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.
[21] A. Löser, C. Tempich, B. Quilitz, W.-T. Balke, S. Staab, and W. Nejdl. Searching dynamic communities with personal indexes. In *International Semantic Web Conference*, pages 491–505, 2005.
[22] J. Lu and J. P. Callan. Content-based retrieval in hybrid peer-to-peer networks. In *CIKM*, pages 199–206, 2003.
[23] T. Luu, F. Klemm, I. Podnar, M. Rajman, and K. Aberer. Alvis peers: a scalable full-text peer-to-peer retrieval engine. In *P2PIR '06: Proceedings of the international workshop on Information retrieval in peer-to-peer networks*, pages 41–48, New York, NY, USA, 2006. ACM Press.
[24] S. Marti, P. Ganesan, and H. Garcia-Molina. DHT routing using social links. In *IPTPS*, pages 100–111, 2004.
[25] W. Meng, C. T. Yu, and K.-L. Liu. Building efficient and effective metasearch engines. *ACM Comput. Surv.*, 34(1):48–89, 2002.
[26] S. Michel, M. Bender, N. Ntarmos, P. Triantafillou, G. Weikum, and C. Zimmer. Discovering and exploiting keyword and attribute-value co-occurrences to improve p2p routing indices. In *CIKM*, pages 172–181, 2006.
[27] S. Michel, M. Bender, P. Triantafillou, and G. Weikum. Iqn routing: Integrating quality and novelty in p2p querying and ranking. In *EDBT*, pages 149–166, 2006.
[28] S. Michel, P. Triantafillou, and G. Weikum. Klee: A framework for distributed top-k query algorithms. In *VLDB*, pages 637–648, 2005.
[29] H. Nottelmann and N. Fuhr. Combining cori and the decision-theoretic approach for advanced resource selection. In *ECIR*, pages 138–153, 2004.
[30] H. Nottelmann and N. Fuhr. Comparing different architectures for query routing in peer-to-peer networks. In *ECIR*, pages 253–264, 2006.
[31] J. X. Parreira, S. Michel, and G. Weikum. p2pdating: Real life inspired semantic overlay networks for web search. *Inf. Process. Manage.*, 43(3):643–664, 2007.
[32] J. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D. H. J. Epema, M. Reinders, M. van Steen, and H. Sips. Tribler: A social-based peer-to-peer system. *Concurrency and Computation: Practice and Experience*, 2007.
[33] A. I. T. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In *Middleware*, pages 329–350, 2001.
[34] L. Si, R. Jin, J. P. Callan, and P. Ogilvie. A language modeling framework for resource selection and results merging. In *CIKM*, pages 391–397, 2002.
[35] K. Sripanidkulchai, B. M. Maggs, and H. Zhang. Efficient content location using interest-based locality in peer-to-peer systems. In *INFOCOM*, 2003.
[36] I. Stoica, R. Morris, D. R. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *SIGCOMM*, pages 149–160, 2001.
[37] C. Tang, Z. Xu, and M. Mahalingam. psearch: information retrieval in structured overlays. *Computer Communication Review*, 33(1):89–94, 2003.
[38] C. Tempich, S. Staab, and A. Wranik. Remindin': semantic query routing in peer-to-peer networks based on social metaphors. In *WWW*, pages 640–649, 2004.