Search Strategies for Scientific Collaboration Networks

Paul - Alexandru Chirita, Andrei Damian, Wolfgang Nejdl, Wolf Siberski
L3S Research Center / University of Hannover
Deutscher Pavillon, Expo Plaza 1
30539 Hannover, Germany
{chirita,damian,nejdl,siberski}@l3s.de

ABSTRACT

Can we improve P2P search by looking into our social network? In this paper, we argue that P2P networks built upon specific communities (e.g., scientific social networks) could achieve such a goal, by providing an implicit personalization to the output results set. Existing work in social networks investigating co-authorship relations has shown that scientific collaboration networks are scale-free. At the same time, P2P systems based on synthesized small-world networks have emerged, with a positive impact on search efficiency. We propose to use existing social collaboration graphs as foundation for the P2P topology instead of creating purely technological topologies. To get an insight into the relationship between scientific collaboration and co-authorship, we compared both for an existing collaboration network. Based on this analysis, we then generated a large P2P collaboration network derived from co-authorship data collections as basis for our experiments. The most prevalent search type in the scientific context is keyword search for relevant publications. We investigate different search strategies suitable in that context and show our initial experimental results.

1. INTRODUCTION

P2P networks are powerful distributed infrastructures capable of handling enormous amounts of resources while maintaining an organized and balanced structure. Yet the more data and peers available, the more efficient algorithms will be needed to accomplish this task. On the other hand, systems (re-)producing social networks have emerged as a very effective way to find persons with related activities, interests, etc. For example, networks like Orkut[1] or Friendster[2] currently already cover several millions of users organized in linked subcommunities.

Can we increase search efficiency and results quality by bringing these two areas together? We claim that the answer is positive, especially when dealing with a scientific environment (i.e., in which authors are writing and searching for articles). First of all, in such an environment a significant part of existing social connections can be automatically inferred (e.g., by analyzing co-authorship relations), and thus the corresponding network topology can be built with little or no manual effort at all. Second, this model also implies a preferential attachment between peers with similar preoccupations. This way, the resources of interest are usually stored in each peer's neighborhood, making it easier to find them. Finally, since the publication is the main item sought for in a scientific environment, we can base query evaluation on existing information retrieval approaches for keyword search (such as TFxIDF or LSI) and get relevant papers by visting only a small number of peers.

In this paper, we investigate several P2P search strategies that exploit the inherent information residing within the social connections between peers in order to enhance both retrieval quality and efficiency. To test these strategies, we need a true model of collaboration relations which are not completely covered by co-authorship relations. Therefore, we start by analyzing collaborations in an existing research group, including those that did not lead to joint peer-reviewed publications. This analysis is presented in Section 2. We apply the characteristics identified in this analysis to the generation of a large simulated collaboration network, and use this as basis for our search experiments. Section 3 discusses possible search strategies and Section 4 presents the outcome of our evaluation. Section 5 discusses related work on social and on P2P networks. Finally, we conclude and discuss further work in Section 6.

2. COLLABORATION NETWORKS

As our research is focused on scientific collaboration networks in which peers share scientific articles, understanding the structure and properties of these networks is very important for developing a realistic model of the P2P infrastructure, as well as for designing efficient search mechanisms to serve it. When modeling collaboration networks, we face the following problem: on the one hand, co-authorship alone is a rather narrow definition for collaboration, since within research groups people often cooperate without publishing a paper together, for instance for writing internal technical papers or seminar presentations that are never published. In addition, co-authorship relations can be crawled only for articles submitted to conferences, workshops and journals that are indexed by the digital libraries currently available (CiteSeer, DBLP, ACM, e-Print Archive, ...). On the other hand, we strive to construct the collaboration network automatically from available data, and that is co-authorship information. In this section we discuss our analysis of collaborations in our research group and describe how the results can be used to extend a co-authorship network in such a way that it approximates the (more connected) collaboration network.

2.1 Co-authorship Networks

Exploiting statistical features of the graph structure for designing

[1] www.orkut.com
[2] www.friendster.com

P2P-IR 2005 Bremen, Germany

efficient network search algorithms has been successfully used not only in the context of the web (see for example [6], or [12]), but also for guided search in P2P networks (see [2], or [1]).

Considering the specifics of the scenario that motivates our research, co-authoship networks are the ideal starting point for the development of our network model. Co-authorship networks offer the largest database to date on social networks and have been the subject of intensive research for understanding the topological and dynamical laws governing complex networks. In [19] and [4], co-authorship graphs for scientists in a variety of fields (computer science, physics, biomedical research, mathematics, neuro-science) are analyzed. All the graphs showed a scale-free character, where the node degree distribution follows a power-law with an exponential cutoff:

$$P(x) = c * x^{-\tau} * e^{-x/z_c} \quad (1)$$

τ and z_c are constants specific to the field of research; for computer science the values are: $\tau = 1.2,\ z_c = 10.7$.

Two other properties making co-authorship networks suitable for our model are the presence of a giant strongly connected component (more than 80% of the nodes are interconnected to each other by paths of intermediate co-authors), and the occurrence of the preferential attachement phenomena, in which nodes link with higher probability to those nodes that already have a large number of links [4].

2.2 Small Collaboration Network Model

Identifying a generic model for scientific *collaboration* is a difficult, if not impossible task, because we can only collect the social relations by specifically requesting them from each person[3]. Therefore, we chose to address this task differently: We built a model of the collaboration network of our group, and then validated this model by manually examining the connections within the EPFL LSIR group.

L3S Collaboration Model: We consider that two people collaborate if they are either co-authors on at least one paper (published or not) or if they weekly exchange articles and scientific references. We have performed an experiment within the L3S research group, on 19 people that included professors, Ph.D. students and undergraduate students, and examined how our definition of collaboration affects the co-authorship model.

We have used the DBLP Computer Science bibliography database to look for the number of co-authors of each test subject and then we asked each of them for an estimation of their number of collaborators according to our definition. We have also asked them for the number of scientific papers they store. The results show that the frequency[4] distribution of the number of co-authors follows a power-law distribution with an exponential cutoff ($y = c * x^{-1.2} * e^{-x/10}$) (as depicted in Figure 1). The power-law coefficient and the cutoff are similar to the results found in literature for computer science co-authorship graphs[5].

The distribution of collaboration links (Figure 2) also follows a power-law distribution with an exponential cutoff, but with a smaller coefficient ($\tau = 0.9$). Basically, the effect of the more comprehensive definition for collaboration is a decrease in the absolute value of the power law exponent, which stands for the fact that the medium connected nodes get significantly more links, while the degree increase for the two extremes (highly connected and weakly connected nodes) is relatively small.

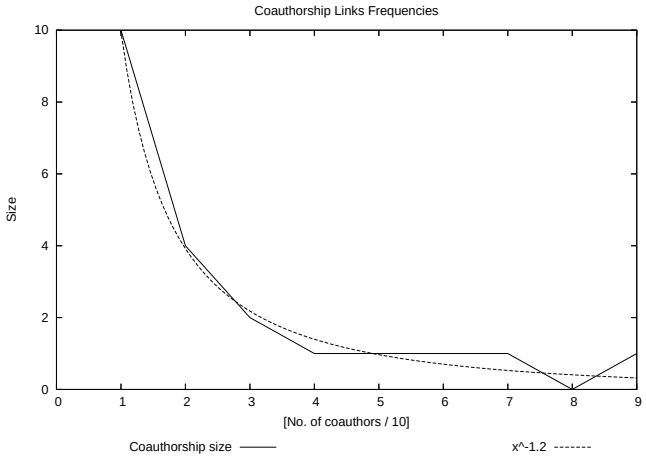

Figure 1: Co-authorship links distribution for the L3S research group

Another parameter we have analyzed was the correlation between the number of collaboration links of each peer and the number of articles she stores on the desktop. Even though there are reasons to believe that peers with many collaborations store many resources as well, by computing the Pearson correlation we found that there is no real association between the two variables:

$$r(X,Y) = \frac{\sum XY - \frac{\sum X \cdot \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N}) \cdot (\sum Y^2 - \frac{(\sum Y)^2}{N})}} = 0.1058 \quad (2)$$

where variable X stands for the number of collaboration links and variable Y stands for the number of resources. By looking at the data, we have noticed two types of user behavior: some peers filter the articles thoroughly and store only the relevant ones, and some others, rather poorly connected, store a large quantity of articles, downloading whole conference proceedings. Locating the latter category is a challenge for the search strategies and a chance to improve search performance as we discuss in Section 3.

2.3 Large Collaboration Network Model

Correlating the results of our analysis of the L3S group with the co-authorship networks properties, we argue that extending a basic co-authorship graph with links such that the degree distribution follows a power law with a smaller absolute coefficient is a valid model for our target network. Since P2P networks are completely connected, we consider only the largest connected component of the co-authorship graph as the starting point for the model and then use preferential attachment for adding collaboration links.

We model preferential attachment by considering the edge clustering coefficient [20]. The edge-clustering coefficient stands for the number of triangles to which a given edge belongs ($z_{i,j}$), divided by the number of triangles that might potentially include it, given the degrees of the adjacent nodes. More formally, for the edge-connecting node i (with degree d_i) to node j (with degree d_j), the edge-clustering coefficient is:

$$C_{i,j} = \frac{z_{i,j} + 1}{min[(d_i - 1), (d_j - 1)]}$$

[3]One could in fact imagine some (semi-)automatic schemes that crawl home pages, such as CiteSeer, but these are also error prone.

[4]Because of the small number of test subjects frequency was computed for intervals of values of size 10.

[5]Using the DBLP co-authorship information, we have also analyzed the EPFL LSIR group and found that it follows the same frequencies distribution as the L3S group.

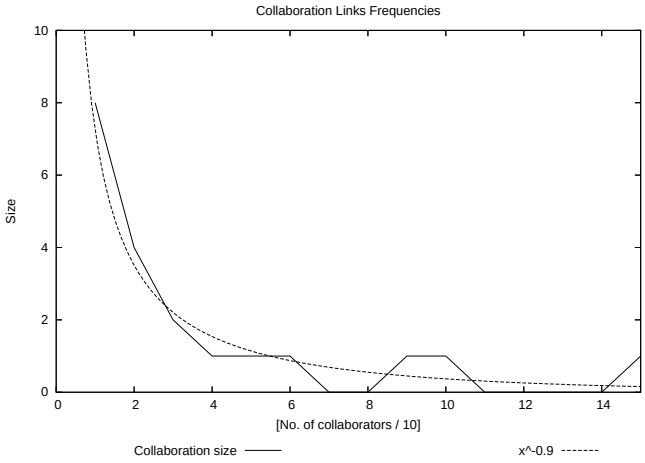

Figure 2: Collaboration links distribution for the L3S research group

For an edge that is not present in the network, this coefficient can be seen as a measure of the effect produced by adding it. The value one is added to the numerator to avoid the degenerate effect caused by zero values for the coefficient. This way, we can add new collaboration edges to the network according to a probability proportional to the clustering coefficient of each edge.

3. EXPLOITING COLLABORATION LINKS FOR P2P SEARCH

By their nature, social networks induce an unstructured topology: A peer is connected to its acquaintances, and all members have equal rights in the community. This model corresponds to the "pure" Peer-to-Peer network architecture. Even though several improvements are possible, we focus on this basic model for the scope of this paper, investigating the options for exploiting the implicit information captured by the social network topology. We will first present a generic framework for searching unstructured P2P networks, and proceed with an analysis of various search strategies, utilizing three different types of information: connectivity, reputation, and peer similarity.

3.1 A Generic Search Framework

We consider unique identification schemes for both peers and queries. Upon issuing a query, a peer p will attach to the query its own ID (or IP address) and a TTL value (equal to the number of peers to visit). It will then send the query to the K most relevant neighbors, following some selection strategy, in this process dividing the TTL value by K.

When a peer receives a query, it r will verify if she has already answered this query. If not, she will start by adding her own ID to the query, then decrease the TTL and locally save the information about the peer from which she received the query. Furthermore, she will check her local index for matches to the query and return the result set to the originating peer[6]. Then she will select the K most relevant neighbors that have not yet been visited and forward the query to them, again splitting the value of the TTL according to the number of peers she forwards the query to. If a peer receives a query she has answered before, she joins the query information with the old value stored locally (i.e., she updates the local information about the source peer ID, as well as the path followed by the query) and then further forward the query. If a peer who receives a query has no unvisited neighbor to forward it, she returns the query to the peer she has received it from. Finally, if the TTL value for a received query has reached zero, the query is not forwarded anymore.

We present the formalized version of our algorithm in pseudo code.

Algorithm 3.1. Generic Search Algorithm for P2P Social Networks.

1: **Basic Data Structures:**
class Peer:
id:int;
index: DocumentIndex;
neighbors: int[];
queriesMap: Map;
Class Query:
id: int;
query: Keyword[];
ttl:int;
visitedPeers:int[];
lastVisited:int;
source:int;

2: **Query processing and forwarding:**
2.1: forwardQuery(Peer p, Query q) :
$selectedPeers$ = selectKMostRelevantPeers(p,q,k);
$expandSize$ = selectedPeers.size();
setLastVisited(q,p);
if ($expandSize > 0$)
splitTTL(q, $expandSize$);
for each r in $selectedPeers$ do receiveQuery(r,q);
else
$querySource$ = get neighbor who submitted the query;
receiveQuery($querySource$, q);
2.2: receiveQuery(Peer r, Query q):
if (not visited(r,q))
addVisited(r,q);
addSourceNeighbour(q,$r.queriesMap$);
decreaseTTL(q);
retrieveLocalResults(r,q);
sendLocalResults(r,q);
if (moreHopsExists(q))
forwardQuery(r,q);
else
if (moreHopsExists(q))
$joinedQuery$ = joinQueryEntries(r,q);
updateLocalMap(r,q,$joinedQuery$);
forwardQuery(r,$joinedQuery$);
2.3: initiateSearch(Peer $initial$, Query q):
setSource($initial$,q);
setInitialTTL(q);
setLastVisited(q,$initial$);
addVisited($initial$,q);
forwardQuery($initial$,q);

[6]An open issue is how to provide peers with reasonable IDF values. A naive approach would be to compute IDFs based on the local document collection only, and use these as basis for scoring. However, this leads to biased scores depending on each peers' documents. Therefore, we must provide all peers with the same IDF values. We see several options to gather this information:

- let some specific peers, e.g., the most highly connected ones, collect document frequencies from their neighborhood and let query issuers ask these peers for the required values;
- use a gossiping approach to gather document frequencies within each highly connected graph component;
- collect and replicate frequency statistics during the query and response distribution.

Each of these approaches has its own advantages and challenges, and we have to investigate which one yields the most accurate estimations, possibly under the least computational effort. For the simulations we provide the peers 'magically' with correct global IDF values. Note that the issue of collection-wide information and the query distribution strategy are independent choices.

3.2 Peer Selection Strategies

Selecting the most relevant peers for query forwarding is a challenging task and involve several questions. First of all and most important, what is relevant? We discuss the relevancy metrics we have implemented in the following subsections, and propose some additional ones in the discussion at the end of Section 4. Second, to how many peers should we send the query to achieve optimal results? We investigate this issue empirically in the next section.

3.2.1 Connectivity-based Selection

The algorithm presented in [2] makes use of the skewed degree distribution of most P2P networks to find the desired results quickly. In their approach the queries are sent only to the best connected neighbor ($K = 1$), while in our search framework we also experiment with different values of K and empirically show that $K = 3$ is optimal. We set the connectivity based selection strategies as a performance baseline for our investigations.

3.2.2 Reputation-based Selection

We have shown in [7] that the previous approach could be slightly improved by selecting the K most *reputable* neighbors, rather than the most connected ones. The reputation measure could be anything from EigenTrust [11] to Distributed Personalized PageRank [7] or some simpler metric [16]. When no personalization is involved, the only advantage is that peers with high quality content, but only few connections, are found faster. On the other hand, a personalized scheme would result in a significant increase in retrieval quality and efficiency. And yet, such a scheme (either personalized or not) needs additional computational effort to generate the reputation values. Therefore, we propose here to alleaviate this problem by introducing the "personalization aspect" into an automatic organization of network topology, based on social relations between people and their membership in different communities. The following subsection discusses a solution which exploits this idea.

3.2.3 Similarity-based Selection

Our model consists of a network generated based on social relations between people. We think it is reasonable to consider that the members within our group (i.e., our co-authors) share similar interests and store more relevant articles than other peers located further away in the social network. Therefore, we investigate in this paper a new peer selection strategy: send the query to our K most "similar" neighbors.

Evaluating the similarity of peers is based on the observation that users' interests are defined by the articles they store on their desktops, in a process of constantly filtering what is important for them. In addition, the stored documents are not isolated resources, but, together with their references, form sparse graphs. For each peer, we define her interest by the graph of stored articles and their references, according to citation relations. We infer the similarity between peers by computing the overlap between their article graphs and propose several measures for computing this metric.

Symmetric Similarity. Let **x** and **y** be vectors of scientific articles together with references stored on the desktops of peers P_1 and P_2 respectively[7]. Then, the *cosine similarity* between P_1 and P_2 can be defined as follows:

$$Cosine(P_1, P_2) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| \cdot |\mathbf{y}|} \quad (3)$$

Likewise, if X and Y are sets of scientific articles, then the cardinal similarity is defined as:

$$Card(P_1, P_2) = \frac{|X \cap Y|}{|X \cup Y|} \quad (4)$$

Relative Similarity. Let **x** and **y** be the above mentioned vectors of scientific articles.

$$Relative(P_1, P_2) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|^2} \quad (5)$$

We think that an asymmetric measure of similarity is more suitable for query forwarding decisions than a symmetric one, since it allows more sensitive decisions. For instance, if peer P_2 has only a subset of the article set of peer P_1, then P_2 will be very interested in sending queries to P_1, while P_1 should not send queries to P_2. We model this scenario by a variation of the $Relative(P_1, P_2)$ measure, and define the *Relative similarity ratio*:

$$RelativeRatio(P_1, P_2) = (\mathbf{x} \cdot \mathbf{y}) \cdot (\frac{|\mathbf{y}|}{|\mathbf{x}|^2}) \quad (6)$$

Similarity based decisions for query forwarding focus search in the community of interest of each peer. Our study gives an empirical evaluation of this strategy, and analyzes the effect of using the different similarity measures proposed above. We show that the *Relative similarity ratio* carries the most useful information for selection decisions.

3.2.4 Hybrid Selection Models

We also investigated the possibilities to combine the similarity and connectivity measures into a hybrid one. Such a scheme considers both orderings of each peer's neighbors, i.e., according similarity and to connectivity values. The algorithm then forwards the query to the K_1 most connected neighbors that have not been visited yet, as well as to the K_2 most similar unvisited ones.

4. SEARCH EVALUATION

4.1 General Setup & Hypotheses

We simulate both a 2000 node and a 10000 node collaboration graph. The nodes in our graphs are authors from the DBLP database[8] and the links represent co-authorship relations. We initiate the graphs starting from Professor Wolfgang Nejdl and Professor Karl Aberer, and then extend following a breadth first iteration over the co-authorship links. For both extensions we have obtained the same $\tau = 1.2$ coefficient for the power law distribution of the nodes' degrees of connectivity.

In order for the co-authorship graphs to fit the proposed collaboration model we have added links considering preferential attachment, such that the degree distribution followed a power law distribution with a coefficient of $\tau = 0.9$ as discussed in Section 2.

As we repeated the same experiment using the CiteSeer OAI metadata, we obtained a slightly modified coefficient for the power-law distribution of the number of co-authors for each node ($\tau = 1.8$). This is due to the fact that while DBLP uses manual annotations and its information is very accurate, CiteSeer uses automatic extraction for article metadata. Because authors may identify themselves in different ways on different papers, e.g., either by using full name, or by using only the initial for the first name, or even by using only initials, the information about each unique author is sometimes split among different nodes. Therefore, the average number of co-authorship relations of each node is smaller, and the

[7]Clearly some peers will not be willing to disclose this information. Then, hashed versions of each paper's title should be used.

[8]We have used the DBLP published XML records.

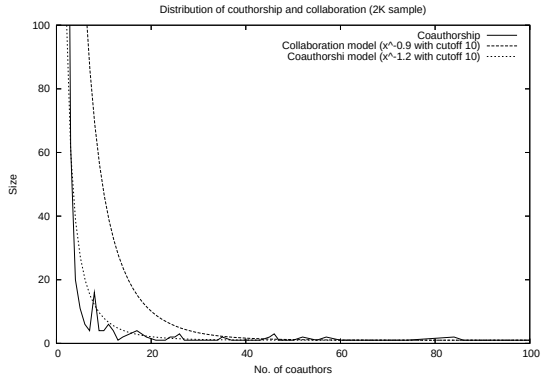

Figure 3: Collaboration links distribution for the 2000 nodes model

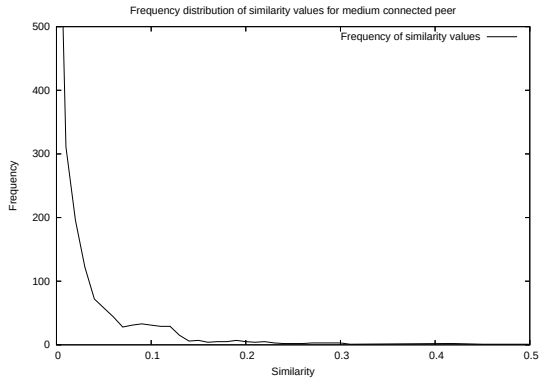

Figure 4: Distribution of the relative similarity values for the 2000 node network

power-law coefficient is greater. However, for our investigations, we used the CiteSeer OAI metadata, as it provides additional meta information besides the title and the authors of each paper (e.g., the abstract, which is not available in the DBLP XML records).

To validate our results from Section 2, we also performed several analyses on this data set. We computed the Pearson correlation between the number of articles of each node, and its number of collaboration links, and we have also found no real association between the two distributions – $r = 0.159$, the result being similar to the one obtained for the real network analyzed in Section 2.2. Both for the L3S group and the generated CiteSeer network, we plotted the relative similarity between (a) a random node and its adjacent nodes, (b) a random node and all other nodes in the network (Figure 4), and (c) all nodes with each other (Figure 5). We found that each of these distributions follows a power-law, thus indicating that each peer has a high similarity only to very few other peers (which presumably also hold the resources deemed interesting by the former peer). This result further motivated our investigation of the similarity based search, since similarity provides an important influence factor for the selection process.

For the experiment we have considered the 2000 node P2P network model discussed above. The resources we used for each peer

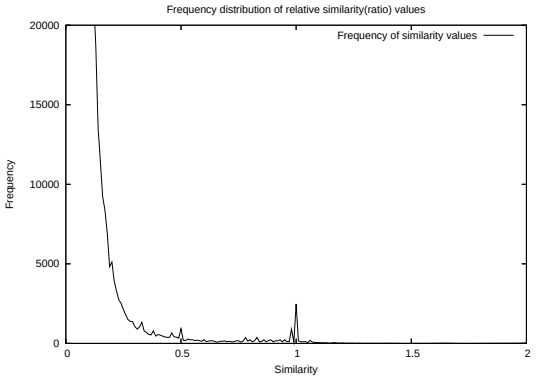

Figure 5: Distribution of the relative similarity values for the 2000 node network

are her authored papers, together with their references. Because of space constraints, for each article we only indexed its title, authors, and abstract[9]. We chose 10 random peers so that their degrees of connectivity are uniformly distributed. For each peer, we automatically generated up to 400 one word and two word queries based on the most frequent terms in their article set[10]. We filtered out queries that returned too many results (more than 5% of the entire article set, because we considered them to be too general) and two word queries that do not respect the constraint:

$$0.01 < \frac{DF_{q_{1,2}}}{min(DF_{q_1}, DF_{q_2})} < 0.1 \quad (7)$$

where q_1 and q_2 are one word queries and $q_{1,2}$ is the combined one. The lower bound stands for the meaningfulness of the association of terms, while the upper bound filters associations that are too common. In the end the union of the query sets had 3600 items.

In the experiments we set the TTL to 100 (5% of the nodes). By further increasing the number of visited nodes, we found that all query forwarding strategies will tend to have similar performance. Moreover, our goal was to get as many best results as possible by visiting only a small percent of nodes (5–10%). Each peer returned its top 50 hits[11], by using TFxIDF with global IDF knowledge (see Footnote 6). As a performance indicator, we used the recall measure, i.e. the degree in which all Top-K matching documents in the collection are returned (from a centralized perspective):

$$Recall = \frac{Number\ of\ TopK\ documents\ returned}{min(K, Number\ of\ matching\ documents)} \quad (8)$$

We investigated the following hypotheses:

1. Similarity based query forwarding decisions (using both the relative similarity and the relative similarity ratio) yield better results in terms of recall than the connectivity based ones.
2. Forwarding queries to the K most similar/most connected neighbors with $K > 3$ will produce only a minimal recall improvement over the strategies using smaller K values.

[9]We have used Lucene for indexing the document sets and an implementation of Porter's algorithm for stemming.

[10]Two word queries were build by combining every two words from the most frequent term list.

[11]We have experimented with the best 10, 20, 30, and 50 results and obtained similar results.

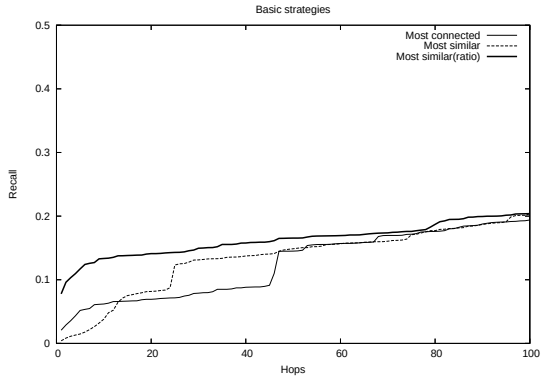

Figure 6: Recall average for basic strategies

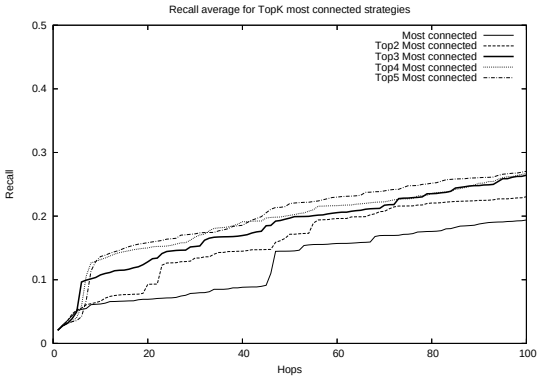

Figure 7: K *most connected* strategies comparison

3. Combining the similarity and connectivity based strategies improves the performance.

Let us now analyze each of these hypotheses in the following subsections.

4.2 Basic strategies for query forwarding

In this experiment we have evaluated the performance of similarity based selection versus connectivity based selection. Each peer forwarded queries only to its best neighbor, according to each strategy. The results (Figure 6) show that by using the relative similarity ratio we obtain the best performance in terms of recall, while using connectivity values or relative similarity values the recall is pretty much the same. This result highlights the problem of using only simple similarity values in ranking the neighboring peers. A very similar neighbor is not always the best place to search for new results, since her article set may not contain too many different items from the current peer's article set. To avoid this problem, the similarity *ratio* should be used instead. Moreover, one also notices that the similarity ratios allow for a good exploration of the source peer's community of interest, as after only 20 visited nodes the results are much better than for the other strategies.

4.3 Top-K strategies for query forwarding

In this experiment each peer forwarded queries to K most relevant of its neighbors, using the basic selection strategies described in the previous experiment. The recall significantly improves over the basic strategies only for $K = 2$ and $K = 3$, whereas for $K > 3$ there is a very low performance increase[12] (Figures 7 and 8). Our experiments also showed that the similarity ratios lead to better heuristics for searching in the peers' community of interest than the simple similarity values or the connectivity information.

4.4 Combined strategies for query forwarding

Motivated by the rather small recall values we also investigated several combined strategies incorporating both the similarity based selection to explore the local communities, as well as the connectivity information to quickly reach distant communities. We expected that in addition to the local community of interest, other nodes holding good results are organized in distant clusters covering the same topic. However, the results from Figures 10 and 9) show that there is no improvement over the best strategies investigated so far, namely "*send to the best 3 peers using the similarity ratios*". This showed that the most connected peers are not always the repository of the best resources in the network.

[12]This can be explained by the power law distribution of the number of connections and of the Peer-to-Peer relative similarity values, since very few peers have many connections / similar neighbors.

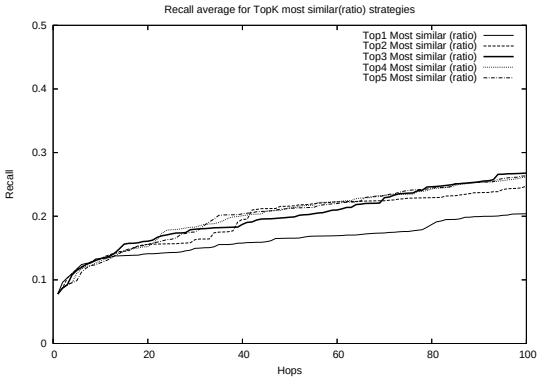

Figure 8: K *most similar (ratio)* strategies comparison

4.5 Discussion

In order to further analyze our results, we plotted the path followed by several queries in the network. One example is presented in Figure 11: the blue nodes represent visited peers that did not provide any Top-K result (we used $K = 50$ here, with the $Top - K$ values calculated from a centralized perspective), the green ones mark peers providing at least one such result, and the red ones denote not visited peers holding a Top-K item. As we can see, the "valuable" peers are rather scattered around the network, and not clustered as we expected from our social model (e.g., peers working on "Information Retrieval" should be somehow closely connected). However, this effect could have also appeared because our automatically generated queries are still rather general.

We are currently investigating a new approach to peer selection, in which the visited peers are selected also according to their ability to answer each specific query (for example, in terms of their document frequency), i.e., in which we combine our query-independent schemes with a query-dependent one.

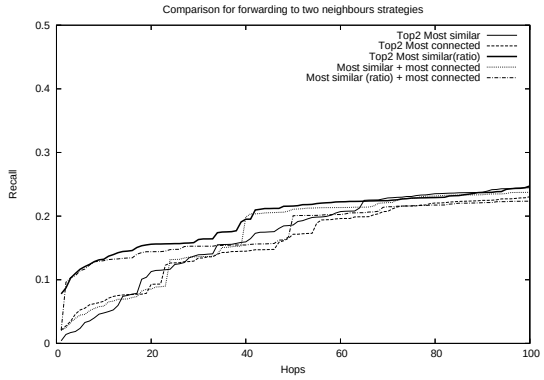

Figure 9: Forward to two neighbors strategies comparison

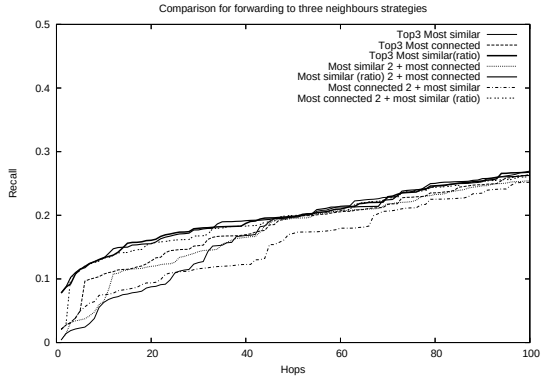

Figure 10: Forward to two neighbors strategies comparison

5. RELATED WORK

Similarity-based network connections. Exploiting peer similarity for improved search is a common topic in P2P approaches. Two different kinds of sources for deriving similarities are used, query results and peer content:

Deriving similarity from query statistics. The main idea here is to continouously optimize connections of each peer based on the responses it gets to its queries. In [21], an importance score is introduced for peers, based on percentage of hits received via this connection, distance and (for direct neighbors) connection time. As soon as an indirectly connected peer becomes more important than a direct one, a direct connection to this peer is established and the connection to the least important immediate neighbor is dropped. [25] score a connection as ratio between the number of times it was successfully used (yielded hits) and the total number of times it was used. They point out that after some time the network topology starts to show small-world characteristics. [31] also do query profiling and select neighbors to forward a query to according to their relevance rank. The relevance ranking does only affect query routing and has no impact on the topology.

Deriving similarity from peer content. This approach is used in P2P information retrieval networks which support keyword queries. SETS [5] is a hybrid topology where the network is split into so-

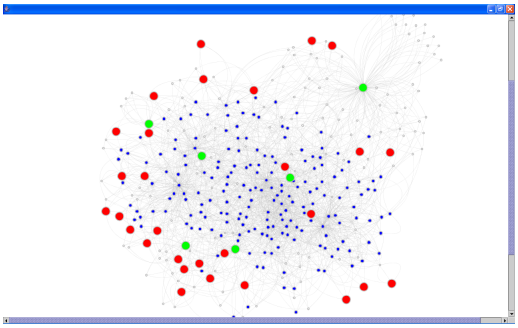

Figure 11: Forward to two neighbors strategies comparison

called topic segments, which arecharacterized by their centroid description. SETS uses a vector space model to represent documents, peers and centroids. A query is first routed to the corresponding segment, and then evaluated using the segment subnet structure. An interesting evaluation result is that using the peer vectors for segmentation outperforms the usage of document vectors as basis for clustering. pSearch [28] sorts peers into a CAN [22] network according to their aggregated latent semantic indexing representation. Schmitz [24] shows how to achieve a small-world topology by organizing peers according to their content. In this approach, a topic classification is used to score similarity. [10] also propose to cluster peers according to their content, but without including a definition for content similarity.

Combining statistics and peer content. In REMINDIN' [29], a connection is scored by the similarity of the query topic to the topic(s) the target peer provides combined with a probability measure that the peer indeed will provide answers. To determine the similarity between query and content, both are annotated using concepts from a shared ontology.

P2P for scientific knowledge exchange. Using P2P to support scientific collaborations has already been suggested in [10]. [18] propose a concrete architecture for Scientific Collaboration Networks (SCN). Both expect this will lead to a small-world topology for the resulting P2P network. OverCite [26] is a project which aims at building a new, P2P-based infrastructure for CiteSeer [13]. Neither of them proposes a peer similarity score based on co-citation or reference links.

Among the related approaches, only Yu and Singh [30] use co-citations to establish connections between peers (agents in their context). As in our case, the experimental context is a simulated network of collaborating scientists. In contrast to their work our peers are not classified using a topic system, and we use another algorithm to increase the assumed collaboration degree. Also, while [30] analyzes how experts can be found within in the network, we investigate the efficiency and effectiveness of queries for content.

Small-world networks for P2P. Many recent systems build on the advantageous characteristics of small-world graphs to improve search efficiency, e.g., [9, 14, 15, 17, 23]. These approaches generate the network using various algorithm, but none of them uses existing social relations as basis for the resulting topology.

P2P information retrieval. P2P networks which evaluate keyword queries using information retrieval algorithms such as TFx-IDF are currently mostly based on structured and/or hierarchical networks (e.g., [8, 27, 28, 3]). To our knowledge, no work has been published regarding the issue of collection-wide information in pure, unstructured networks. The only P2P IR approach based on an unstructured network we are aware of just ignores this issue and computes scores based on local peer collections [31].

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented several strategies for searching scientific collaboration P2P networks by exploiting the implicit social organization provided by this model. We have empirically shown that similarity-based search strategies have better performance than the connectivity based ones. We have also found that selecting more than three peers at each step will not increase overall performance. As future work, we intend to integrate the use of search performance history into our search framework and study the performance gain under each technique.

7. REFERENCES

[1] L.A. Adamic, R. Lukose, and B. Huberman. Local search in unstructured networks. In *Handbook of Graphs and Networks*, pages 295–317. Wiley-VCH, 2003.

[2] L.A. Adamic, R. Lukose, A. Puniyani, and B. Huberman. Search in power-law networks. *Phys. Rev.*, E 64, 046135, 2001.

[3] Wolf-Tilo Balke, Wolfgang Nejdl, Wolf Siberski, and Uwe Thaden. Progressive distributed top-k retrieval in peer-to-peer networks. In *IEEE Intl. Conf. on Data Engineering (ICDE)*, 2005.

[4] A. L. Barabasi, H. Jeong, E. Ravasz, Z. Neda, A. Schuberts, and T. Vicsek. Evolution of the social network of scientific collaborations. In *Physica A 311*, pages 590–614, 2002.

[5] M. Bawa, G. S. Manku, and P. Raghavan. Sets: search enhanced by topic segmentation. In *Proc. of the 26th Annual Intl. ACM SIGIR Conference*, Toronto, Canada, 2003.

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[7] Paul-Alexandru Chirita, Wolfgang Nejdl, and Oana Scurtu. Knowing where to search: Personalized search strategies for peers in p2p networks. In *Proc. of the Peer-to-Peer Information Retrieval Workshop held at the 27th Intl. ACM SIGIR Conference*, 2004.

[8] F. Cuenca-Acuna, C. Peery, R. Martin, and T. Nguyen. Planetp: Using gossiping to build content addressable peer-to-peer information sharing communities. In *IEEE Intl. Symp. on High Performance Distributed Computing*, 2003.

[9] K. Hui, J. Lui, and D. Yau. Small world overlay p2p networks. In *Proc. of 12th Intl. Workshop on Quality of Service (IWQoS 2004)*, Montreal, Canada, 2004.

[10] Adriana Iamnitchi, Matei Ripeanu, and Ian T. Foster. Locating data in (small-world?) peer-to-peer scientific collaborations. In *Proc. of the First Intl. Workshop on Peer-to-Peer Systems*, Cambridge, MA, USA, 2002.

[11] S. Kamvar, M. Schlosser, and H. Garcia-Molina. The EigenTrust Algorithm for Reputation Management in P2P Networks. In *Proc. of the 12th Intl. WWW Conference*, 2003.

[12] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[13] S. Lawrence, C. L. Giles, and K. D. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

[14] Mei Li, Wang-Chien Lee, and Anand Sivasubramaniam. Semantic small world: An overlay network for peer-to-peer search. In *Proc. of the 12th IEEE Intl. Conf. on Network Protocols*, Berlin, Germany, 2004.

[15] G Manku, M Bawa, and P Raghavan. Symphony: Distributed hashing in a small world. In *Proceedings of the 4th USENIX Symposium on Internet Technologies and Systems*, 2003.

[16] S. Marti and H. Garcia-Molina. Limited Reputation Sharing in P2P Systems. In *Proceedings of ACM Conference on Electronic Commerce (EC04)*, 2004.

[17] S. Merugu, S. Srinivasan, and E. W. Zegura. Adding structure to unstructured peer-to-peer networks: the use of small-world graphs. *Journal of Parallel Distrib. Comput.*, 65(2):142–153, 2005.

[18] J. Mitre and L. Navarro-Moldes. P2p architecture for scientific collaboration. In *Proc. of the 14th IEEE Intl. Workshop on Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE-2004)*, 2004.

[19] M.E.J. Newman. The structure of scientific collaboration networks. In *Proc. Natl. Acad. Sci. USA 98*.

[20] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. In *Preprint cond-mat/0309488*, 2003.

[21] Murali Krishna Ramanathan, Vana Kalogeraki, and Jim Pruyne. Finding good peers in peer-to-peer networks. In *Proc. of the Intl. Parallel and Distributed Processing Symposium*, Fort Lauderdale, USA, 2002.

[22] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker. A scalable content-addressable network. In *Proc. of the 2001 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Comm.*, 2001.

[23] N. Sarshar, P. O. Boykin, and V. P. Roychowdhury. Percolation search in power law networks: Making unstructured peer-to-peer networks scalable. In *Proc. of the 4th Intl. Conf. on Peer-to-Peer Computing*, Zurich, Switzerland, 2004.

[24] Christoph Schmitz. Self-organization of a small world by topic. In *Proc. of the 1st Intl. Workshop on Peer-to-Peer Knowledge Management*, Boston, MA, USA, 2004.

[25] K. Sripanidkulchai, B. M. Maggs, and H. Zhang. Efficient content location using interest-based locality in peer-to-peer systems. In *Proc. of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM2003)*, San Franciso, USA, 2003.

[26] J. Stribling, I. Councill, J. Li, M. F. Kaashoek, D. R. Karger, R. Morris, and S. Shenker. Overcite: A cooperative digital research library. In *Proc. of the 4th Intl. Workshop on Peer-To-Peer Systems (IPTPS'05)*, 2005.

[27] T. Suel, C. Mathur, J. Wu, J. Zhang, A. Delis, M. Kharrazi, X. Long, and K. Shanmugasundaram. Odissea: A peer-to-peer architecture for scalable web search and information retrieval. In *Proc. of the Intl. Workshop on Web and Databases*, 2003.

[28] Chunqiang Tang, Sandhya Dwarkadas, and Zhichen Xu. On scaling latent semantic indexing for large peer-to-peer systems. In *Proc. of the 27th Annual Intl. ACM SIGIR Conference*, Sheffield, UK, 2004.

[29] Christoph Tempich, Steffen Staab, and Adrian Wranik. REMINDIN': Semantic query routing in peer-to-peer networks based on social metaphors. In *Proceedings of the 13th Intl. WWW Conference*, 2004.

[30] Bin Yu and Munindar P. Singh. Searching social networks. In *Proc. of the Second Intl. Joint Conference on Autonomous Agents & Multiagent Systems (AAMAS 2003)*, 2003.

[31] D. Zeinalipour-Yazti, V. Kalogeraki, and D. Gunopulos. Exploiting locality for scalable information retrieval in peer-to-peer networks. *Inf. Syst.*, 30(4):277–298, 2005.