Peer Selection in Peer-to-Peer Networks with Semantic Topologies

Peter Haase[1], Ronny Siebes[2], and Frank van Harmelen[2]

[1] Institute AIFB, University of Karlsruhe, D-76128 Karlsruhe, Germany
haase@aifb.uni-karlsruhe.de
[2] Vrije Universiteit Amsterdam, The Netherlands
{ronny,frankh}@cs.vu.nl

Abstract. Peer-to-Peer systems have proven to be an effective way of sharing data. Modern protocols are able to efficiently route a message to a given peer. However, determining the destination peer in the first place is not always trivial. We propose a model in which peers advertise their expertise in the Peer-to-Peer network. The knowledge about the expertise of other peers forms a semantic topology. Based on the semantic similarity between the subject of a query and the expertise of other peers, a peer can select appropriate peers to forward queries to, instead of broadcasting the query or sending it to a random set of peers. To calculate our semantic similarity measure we make the simplifying assumption that the peers share the same ontology. We evaluate the model in a bibliographic scenario, where peers share bibliographic descriptions of publications among each other. In simulation experiments we show how expertise based peer selection improves the performance of a Peer-to-Peer system with respect to precision, recall and the number of messages.

1 Introduction

Peer-to-Peer systems are distributed systems without any centralized control or hierarchical organization, in which each node runs software with equivalent functionality. A review of the features of recent Peer-to-Peer applications yields a long list: redundant storage, permanence, selection of nearby servers, anonymity, search, authentication, and hierarchical naming. Despite this rich set of features, scalability is a significant challenge: Peer-to-Peer networks that broadcast all queries to all peers don't scale - intelligent query routing and network topologies are required to be able to route queries to a relevant subset of peers. Modern routing protocols like Chord [15], CAN [14] are based on the idea of Distributed Hash Tables for efficient query routing, but little effort has been made with respect to rich semantic representations of metadata and query functionalities beyond simple keyword searches.

The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation [2]. In a distributed knowledge management system these Semantic Web techniques can be used for expressing the knowledge shared by peers in a well-defined and formal way.

In the model that we propose, peers use a shared ontology to advertise their expertise in the Peer-to-Peer network. The knowledge about the expertise of other peers forms a

M. Bouzeghoub et al. (Eds.): ICSNW 2004, LNCS 3226, pp. 108–125, 2004.

semantic topology, independent of the underlying network topology. If the peer receives a query, it can decide to forward it to peers about which it knows that their expertise is *similar* to the subject of the query. The advantage of this approach is that queries will not be forwarded to all or a random set of known peers, but only to those that have a good chance of answering it.

In this paper we instantiate the above model with a bibliographic scenario, in which researchers share bibliographic metadata about publications. In the evaluation of our model we will show how

- the proposed model of expertise based peer selection considerably improves the performance of the Peer-to-Peer system,
- ontology-based matching with a similarity measure will improve the system compared with an approach that relies on exact matches, such as a simple keyword based approach,
- the performance of the system can be improved further, if the semantic topology is built according to the semantic similarity of the expertises of the peers,
- a "perfect" semantic topology imposed on the network using global knowledge yields ideal results.

In the remainder of the paper we will present the formal model for expertise base peer selection (Section 2), instantiate this model for the bibliographic scenario (Section 3), define evaluation criteria (Section 4), present results of the simulation (Section 5), discuss related work (Section 6) and conclude with some directions for future work (Section 7).

2 A Model for Expertise Based Peer Selection

In the model we propose, peers advertise their expertise in the network. The peer selection is based on matching the subject of a query and the expertise according to their semantic similarity. Figure 1 below shows the idea of the model in one picture.

In this section we first introduce a model to semantically describe the expertise of peers and how peers promote their expertise as advertisement messages in the network. Second, we describe how the received advertisements allows a peer to select other peers

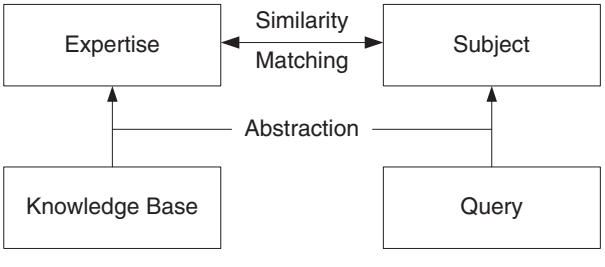

Fig. 1. Expertise Based Matching

for a given query based on a semantic matching of query subjects against expertise descriptions. The third part describes how a *semantic topology* can be formed by advertising expertise.

2.1 Semantic Description of Expertise

Peers. The Peer-to-Peer network consists of a set of peers P. Every peer $p \in P$ has a knowledge base that contains the knowledge that it wants to share.

Common Ontology. The peers share an ontology O, which provides a common conceptualization of their domain. The ontology is used for describing the expertise of peers and the subject of queries.

Expertise. An expertise description $e \in E$ is a abstract, semantic description of the knowledge base of a peer based on the common ontology O. This expertise can either be extracted from the knowledge base automatically or specified in some other manner.

Advertisements. Advertisements $A \subseteq P \times E$ are used to promote descriptions of the expertise of peers in the network. An advertisement $a \in A$ associates a peer p with a an expertise e. Peers decide autonomously, without central control, whom to promote advertisements to and which advertisements to accept. This decision can be based on the semantic similarity between expertise descriptions.

2.2 Matching and Peer Selection

Queries. Queries $q \in Q$ are posed by a user and are evaluated against the knowledge bases of the peers. First a peer evaluates the query against its local knowledge base and then decides which peers the query should be forwarded to. Query results are returned to the peer that originally initiated the query.

Subjects. A subject $s \in S$ is an abstraction of a given query q expressed in terms of the common ontology. The subject can be seen a complement to an expertise description, as it specifies the required expertise to answer the query.

Similarity Function. The similarity function $SF : S \times E \mapsto [0, 1]$ yields the semantic similarity between a subject $s \in S$ and an expertise description $e \in E$. An increasing value indicates increasing similarity. If the value is 0, s and e are not similar at all, if the value is 1, they match exactly. SF is used for determining to which peers a query should be forwarded. Analogously, a same kind of similarity function $E \times E \mapsto [0, 1]$ can be defined to determine the similarity between the expertise of two peers.

Peer Selection Algorithm. The peer selection algorithm returns a ranked set of peers. The rank value is equal to the similarity value provided by the similarity function.

From this set of ranked peers one can, for example, select the best n peers, or all peers whose rank value is above a certain threshold, etc.

Algorithm 1 Peer Selection

let A be the advertisements that are available on the peer
let γ be the minimal similarity between the expertise of a peer and the topics of the query.
$subject := ExtractSubject(query)$
$rankedPeers := \emptyset$
for all $ad \in A$ **do**
  $peer := Peer(ad)$
  $rank := SF(Expertise(ad), subject)$
  if $rank > \gamma$ **then**
    $rankedPeers := (peer, rank) \cup rankedPeers$
return $rankedPeers$

2.3 Semantic Topology

The knowledge of the peers about the expertise of other peers is the basis for a semantic topology. Here it is important to state that this semantic topology is independent of the underlying network topology. At this point, we don't make any assumptions about the properties of the topology on the network layer.
The semantic topology can be described by the following relation:

$Knows \subseteq P \times P$, where $Knows(p_1, p_2)$ means that p_1 knows about the expertise of p_2.

The relation $Knows$ is established by the selection of which peers a peer sends its advertisements to. Furthermore, peers can decide to accept an advertisement, e.g. to include it in their registries, or to discard the advertisement. The semantic topology in combination with the expertise based peer selection is the basis for intelligent query routing.

3 The Bibliographic Scenario

In this section we instantiate the general model for expertise based peer selection from previous section. We use a real-life scenario for knowledge sharing in a Peer-to-Peer environment.

In the daily life of a computer scientist, one regularly has to search for publications or their correct bibliographic metadata. Currently, people do these searches with search engines like Google and CiteSeer, via university libraries or by simply asking other people that are likely to know how to obtain the desired information.

The scenario that we envision here is that researchers in a community share bibliographic metadata via a Peer-to-Peer system. The data may have been obtained from BibTeX files or from a bibliography server such as the DBLP database[1]. A similar scenario is described in [1], where data providers, i.e. research institutes, form a Peer-to-Peer network which supports distributed search over all the connected metadata repositories.

We now describe the bibliographic scenario using the general model presented in the previous section.

[1] http://dblp.uni-trier.de/

Peers. A researcher is represented by a peer $p \in P$. Each peer has an RDF knowledge base, which consists of a set of bibliographic metadata items that are classified according to the ACM topic hierarchy[2]. The following example shows a fragment of a sample bibliographic item based on the Semantic Web Research Community Ontology (SWRC)[3]:

```
<rdf:RDF xmlns=
 "http://www.semanticweb.org/ontologies/swrc-onto.daml#"
  xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:acm ="http://daml.umbc.edu/ontologies/topic-ont#">
<Publication rdf:about="dblp:persons/Codd81">
 <title>The Capabilities of
        Relational Database Management Systems.</title>
 <acm:topic rdf:resource=
   "http://daml.umbc.edu/ontologies/classification#
    ACMTopic/Information_Systems/Database_Management"/>
 <!-- ... -->
</Publication>
</rdf:RDF>
```

Common Ontology. The ontology O that is shared by all the peers is the ACM topic hierarchy. The topic hierarchy contains a set, T, of 1287 topics in the computer science domain and relations $(T \times T)$ between them: *SubTopic* and *seeAlso*.

Expertise. The ACM topic hierarchy is the basis for our expertise model. Expertise E is defined as $E \subseteq 2^T$, where each $e \in E$ denotes a set of ACM topics, for which a peer provides classified instances.

Advertisements. Advertisements associate peers with their expertise: $A \subseteq P \times E$. A single advertisement therefore consists of a set of ACM topics for which the peer is an expert on.

Queries. We use the RDF query language SeRQL [6] to express queries against the RDF knowledge base of a peer. The following sample query asks for publications with their title about the ACM topic *Information Systems / Database Management*:

```
CONSTRUCT {pub} <swrc:title> {title} FROM
{Subject} <rdf:type> {<swrc:Publication>};
  <swrc:title> {title};
  <acm:topic>
  {<topic:ACMTopic/Information_Systems/Database_Management>}
USING NAMESPACE
swrc=<!http://www.semanticweb.org/ontologies/swrc-onto.daml#>,
rdf =<!http://www.w3.org/1999/02/22-rdf-syntax-ns#>,
acm =<!http://daml.umbc.edu/ontologies/topic-ont#>,
topic=<!http://daml.umbc.edu/ontologies/classification#>
```

[2] http://www.cs.vu.nl/ heiner/public/SW@VU/classification.daml
[3] http://ontobroker.semanticweb.org/ontos/swrc.html

Subjects. Analogously to the expertise, a subject $s \in S$ is an abstraction of a query q. In our scenario, each s is a set of ACM topics, thus $s \subseteq T$. For example, the extracted subject of the query above would be *Information Systems/Database Management*.

Similarity Function. In this scenario, the similarity function SF is based on the idea that topics which are close according to their positions in the topic hierarchy are more similar than topics that have a larger distance. For example, an expert on ACM topic *Information Systems/Information Storage and Retrieval* has a higher chance of giving a correct answer on a query about *Information Systems/Database Management* than an expert on a less similar topic like *Hardware/Memory Structures*.

To be able to define the similarity of a peer's expertise and a query subject, which are both represented as a set of topics, we first define the similarity for individual topics. [10] have compared different similarity measures and have shown that for measuring the similarity between concepts in a hierarchically structured semantic network, like the ACM topic hierarchy, the following similarity measure yields the best results:

$$S(t_1, t_2) = \begin{cases} e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} & \text{if } t_1 \neq t_2, \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

Here l is the length of the shortest path between topic t_1 and t_2 in the graph spanned by the *SubTopic* relation. h is the level in the tree of the direct common subsumer from t_1 and t_2.

$\alpha \geq 0$ and $\beta \geq 0$ are parameters scaling the contribution of shortest path length l and depth h, respectively. Based on their benchmark data set, the optimal values are: $\alpha = 0.2, \beta = 0.6$. Using the shortest path between two topics is a measure for similarity because Rada et al [13] have proven that the minimum number of edges separating topics t_1 and t_2 is a metric for measuring the conceptual distance of t_1 and t_2. The intuition behind using the depth of the direct common subsumer in the calculation is that topics at upper layers of hierarchical semantic nets are more general and are semantically less similar than topics at lower levels.

Now that we have a function for calculating the similarity between two individual topics, we define SF as:

$$SF(s, e) = \frac{1}{|s|} \sum_{t_i \in s} \max_{t_j \in e} S(t_i, t_j) \tag{2}$$

With this function we iterate over all topics of the subject and average their similarities with the most similar topic of the expertise.

Peer Selection Algorithm. The peer selection algorithm ranks the known peers according to the similarity function described above. Therefore, peers that have an expertise more similar to that of the subject of the query will have a higher rank. From the set of ranked peers, we now only consider a selection algorithm that selects the best n peers.

4 Evaluation Criteria

In this section we define a number of criteria for a Peer-to-Peer system, which will be the basis for the evaluation of our proposed model for peer selection. These criteria are mainly based on those described in [7].

4.1 Input Parameters

The following input parameters are important criteria that influence the performance of a Peer-to-Peer system:

Number of Peers. The size of the Peer-to-Peer network is represented by this number. Typically the scalability of the system is measured in terms of number of peers.

Number of Documents. The scalability of a Peer-to-Peer system can also be expressed in terms of the number of shared resource items, e.g. documents.

Document Distribution. The document distribution in Peer-to-Peer networks is rarely completely random, but often has certain properties. With this input parameter we want to evaluate how the proposed model behaves with different document distributions.

Network Topology. The performance of a Peer-to-Peer system is strongly influenced by the network topology and its characteristics. Possible topologies could for example be super-peer based, star or ring-shaped, or simply a random graph.

Advertisements. The advertisements are responsible for building the semantic topology. There are various variables involved, e.g. whom to send the advertisments to and which received advertisements to include based on the semantic similarity between the own expertise and that of the advertisement.

Peer Selection Algorithm. The peer selection algorithm determines which peers a query should be forwarded to. This could be a naive algorithm, which simply broadcasts a query, or a more advanced one, as the proposed expertise based peer selection.

Maximum Number of Hops. The maximum number of hops determines how many times a query is allowed to be forwarded. It determines how much the network will be flooded by a single query.

4.2 Output Parameters

To evaluate a Peer-to-Peer system, we use precision and recall measures known from classical Information Retrieval. Here we distinguish measures on the document level (query answering) and the peer level (peer selection). These measures are defined as follows:

Document level (Query Answering)

$Precision_{Doc} = \frac{|A \bigcap B|}{|B|}$
: indicates how many of the returned documents are relevant, with A being the set of relevant documents in the network and B being the set of returned documents. In our model we work with exact queries, therefore only relevant documents are returned. The precision will therefore always be one:
$Precision_{Doc} = \frac{|B|}{|B|} = 1.$

$Recall_{Inf} = \frac{|A \bigcap B|}{|A|} = \frac{|B|}{|A|}$
: The recall on the document level states how many of the relevant documents are returned.

Peer Level (Peer Selection)

$Precision_{Peer} = \frac{|A \bigcap B|}{|B|}$
: For a given query, how many of the peers that were selected had relevant information. Here A is the set of peers that had relevant documents and B is the set of peers that were reached.

$Recall_{Peer} = \frac{|A \bigcap B|}{|A|}$
: indicates for a given query, how many of the peers that had relevant information were reached.

Further Parameters. Another important output parameters is:

$Number_{Messages}$
: This output parameter indicates with how many messages the network is flooded by one query. The number of messages does not only affect the network traffic, but also CPU consumption, such as for the processing of the queries in the case of query messages.

Other output parameters that might be used as evaluation criteria, but are not considered in the following, are for example the size of messages and response times, as they are not relevant for the evaluation of our model.

5 Experimental Results

In this section we describe the simulation of the scenario presented in section 3. The evaluations are based on the criteria defined in section 4. With the experiments we try to validate the following hypotheses:

- **H1 - Expertise based selection:** The proposed approach of expertise based peer selection yields better results than a naive approach based on random selection. The higher precision of the expertise based selection results in a higher recall of peers and documents, while reducing the number of messages per query.
- **H2 - Ontology based matching:** Using a shared ontology with a metric for semantic similarity improves the recall rate of the system compared with an approach that relies on exact matches, such as a simple keyword based approach.

- **H3 - Semantic topology:** The performance of the system can be improved further, if the semantic topology is built according to the semantic similarity of the expertises of the peers. This can be realized, for example, by accepting advertisements that are semantically similar to the own expertise.
- **H4 - The "Perfect" topology:** Perfect results in terms of precision and recall can be achieved, if the semantic topology coincides with a distribution of the documents according to the expertise model.

Data Set. To obtain a critical mass of bibliographic data, we used the DBLP data set, which consists of metadata for 380440 publications in the computer science domain.

We have classified the publications of the DBLP data set according to the ACM topic hierarchy using a simple classification scheme based on lexical analysis: A publication is said to be about a topic, if the label of the topic occurs in the title of the publication. For example, a publication with the title "The Capabilities of Relational Database Management Systems." is classified into the topic *Database Management*.Topics with labels that are not unique (e.g. *General* is a subtopic of both *General Literature* and *Hardware*) have been excluded from the classification, because typically these labels are too general and would result in publications classified into multiple, distant topics in the hierarchy. Obviously, this method of classification is not as precise as a sophisticated or manual classification. However, a high precision of the classification is not required for the purpose of our simulations. As a result of the classification, about one third of the DBLP publications (126247 out of 380440) have been classified, where 553 out of the 1287 ACM topics actually have classified publications. The classified DBLP subset has been used for our simulations.

Document Distribution. We have simulated and evaluated the scenario with two different distributions, which we describe in the following. Note that for the simulation of the scenario we disregard the actual documents and only distribute the bibliographic metadata of the publications.

Topic Distribution: In the first distribution, the bibliographic metadata are distributed according to their topic classification. There is one dedicated peer for each of the 1287 ACM topics. The distribution is directly correlated with the expertise model, each peer is an expert on exactly one ACM topic and contains all the corresponding publications. This also implies that there are peers that do not contain publications, because not all topics have classified instances.

Proceedings Distribution: In the second distribution, the bibliographic metadata are distributed according to conference proceedings and journals in which the according publications were published. For each of the conference proceedings and journals covered in DBLP there is a dedicated peer that contains all the associated publication descriptions (in the case of the 328 journals) or inproceedings (in the case of the 2006 conference proceedings). Publications that are published neither in a journal nor in conference proceedings are contained by one separate peer. The total number of peers therefore is 2335 (=328+2006+1). With this distribution one peer can be an expert on multiple topics, as a journal or conference typically covers mutliple ACM topics. Note that there is still a correlation between the distribution and the expertise, as a conference or journal typically covers a coherent set of topics.

Simulation Environment. To simulate the scenario we have developed and used a controlled, configurable Peer-to-Peer simulation environment. A single simulation experiment consists of the following sequence of operations:

1. *Setup network topology:* In the first step we create the peers with their knowledge bases according to the document distribution and arrange them in a random network topology, where every peer knows 10 random peers. We do not make any further assumptions about the network topology.
2. *Advertising Knowledge:* In the second step, the semantic topology is created. Every peer sends an advertisement of its expertise to all other peers it knows based on the network topology. When a peer receives an advertisement, it may decide to store all or selected advertisements, e.g. if the advertised expertise is semantically similar to its own expertise. After this step the semantic topology is static and will not change anymore.
3. *Query Processing:* The peers randomly initiate queries from a set of randomly created 12870 queries, 10 for each of the 1287 ACM topic. The peers first evaluate the queries against their local knowledge base and then propagate the query according to their peer selection algorithms described below.

Experimental Settings. In our experiments we have systematically simulated various settings with different values of input variables. In the following we will describe an interesting selected subset of the settings to prove the validity of our hypotheses.

Setting 1. In the first setting we use a naive peer selection algorithm, which selects n *random* peers from the set of peers that are known from advertisements received, but disregarding the content of the advertisement. In the experiments, we have used n=2 in every setting, as a rather arbitrary choice.

Setting 2. In the second setting we apply the expertise based selection algorithm. The *best* n (n=2) peers are selected for query forwarding. Here the peer selection algorithm only considers *exact* matches of topics.

Setting 3. In the third setting we modify the peer selection algorithm to use the ontology based similarity measure, instead of only exact matches. The peer selection only selects peers whose expertise is equally or more similar to the subject of the query than the expertise of the forwarding peer.

Setting 4. In the fourth setting we modify the peer to only accept advertisements that are semantically similar to its own expertise. The threshold for accepting advertisements was set to accept on average half of the incoming advertisements.

Setting 5. In this setting we assume global knowledge to impose a perfect topology on the peer network. In this perfect topology the *knows* relation conincides with the ACM topic hierarchy: Every peer knows exactly those peers that are experts on the neighboring topics of its own expertise. This setting is only applicable for the distribution of the publications according to their topics, as this model assumes exactly one expert per topic.

The following table summarizes the instantiations of the input variables for the described settings:

Setting #	Peer Selection	Advertisements	Topology
Setting 1	random	accept all	random
Setting 2	exact match	accept all	random
Setting 3	ontology based match	accept all	random
Setting 4	ontology based match	accept similar	random
Setting 5	ontology based match	accept similar	perfect

Simulation Results. Figures 2 through 5 show the results for the different settings and distributions. The simulations have been run with a varying number of allowed hops. In the results we show the performance for a maximum of up to eight hops. Zero hops means that the query is processed locally and not forwarded. Please note that the diagrams for the number of messages per query and recall (i.e. Figures 5, 3, 4) present cumulative values, i.e. they include the sum of the results for *up to* n hops. The diagram for the precision (Figure 2) of the peer selection displays the precision for a particular number of hops.

In the following, we will interpret the results of the experiments for the various settings described above with respect to our hypotheses H1 through H4.

R1 - Expertise based selection. The results of Figure 2, Setting 1, show that the naive approach of random peer selection gives a constant low precision of 0.03% for the topic distribution and 1.3% for the proceedings distribution. This results in a fairly low recall of peers and documents despite a high number of messages, as shown in Figures 3, 5, 4, respectively. With the expertise based selection, either exact or similarity based matching, the precision can be improved considerably by about one order of magnitude. For example, with the expertise based selection in Setting 3, the precision of the peer selection (Figure 2) can be improved from 0.03% to 0.15% for the topic distribution and from 1.3% to 15% for the proceedings distribution. With the precision, also the recall of peers and documents rises (Figures 3, 5). At the same time, the number of messages per

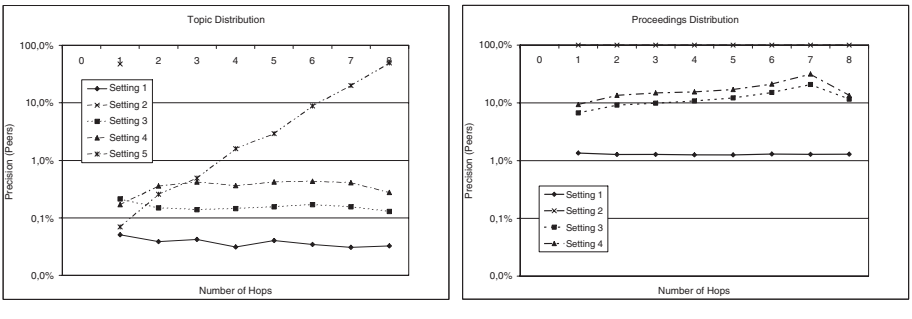

Fig. 2. $Precision_{Peers}$

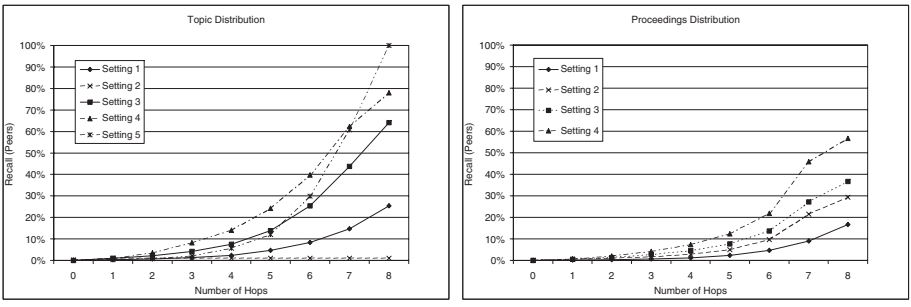

Fig. 3. $Recall_{Peers}$

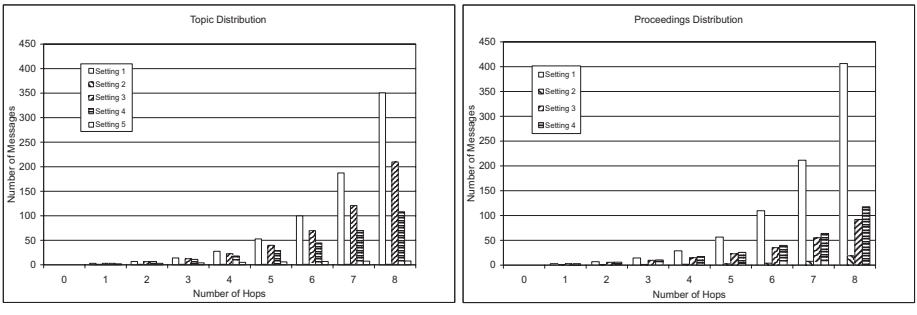

Fig. 4. $Number_{Messages}$

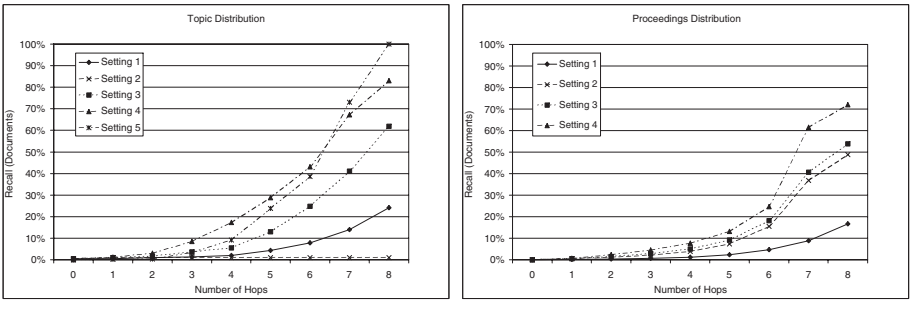

Fig. 5. $Recall_{Documents}$

query can be reduced. The number of messages sent is influenced by two effects. The first effect is message redundancy: The more precise the peer selection, the higher is the chance of a peer receiving a query multiple times on different routes. This redundancy is detected by the receiving peer, which will forward the query only once, thus resulting in a decreasing number of queries sent across the network. The other effect is caused by the selectivity of the peer selection: It only forwards the query to peers whose expertise is semantically more or equally similar to the query than that of the own expertise. With

an increasing number of hops, as the semantic similarity of the expertise of the peer and the query increases, the chance of knowing a qualifying peer decreases, which results in a decrease of messages.

R2 - Ontology based matching. The result of Figure 2, Setting 2, shows that the exact match approach results in a maximum precision already after one hop, which is obvious because it only selects peers that match exactly with the query's subject. However, Figure 3 shows that the recall in this case is very low in the case of the topic distribution. This can be explained as follows: For every query subject, there is only one peer that exactly matches in the entire network. In a sparse topology, the chance of knowing that relevant peer is very low. Thus the query cannot spread effectively across the network, resulting in a document recall of only 1%. In contrary, Setting 3 shows that when semantically similar peers are selected, it is possible to improve the recall of peers and documents, to 62% after eight hops. Also in the case of the proceedings distribution, where multiple exact matches are possible, we see an improvement from 49% in the case of exact matches (Setting 2), to 54% in the case of ontology based matches (Setting 3). Naturally, this approach requires to send more messages per query and also results in a lower precision.

R3 - Semantic Topology. In Setting 4 the peers only accept semantically similar advertisements. This has proven to be a simple, but effective way for creating a semantic topology that correlates with the expertise of the peers. This allows to forward queries along the gradient of increasing semantic similarity. When we compare this approach with that of Setting 3, the precision of the peer selection can be improved from 0.15% to 0.4% for the topic distribution and from 14% to 20% for the proceedings distribution. The recall of documents can thus be improved from 62% to 83% for the topic distribution and from 54% to 72% for the proceedings distribution.

It is also interesting to note that the precision of the peer selection for the similarity based matching decreases slightly after seven hops (Figure 2). The reason is that after seven hops the majority of the relevant peers has already been reached. Thus the chance of finding relevant peers decreases, resulting in a lower precision of the peer selection.

R4 - The "Perfect" Topology. The results for Setting 5 show how one could obtain the maximum recall and precision, if it were possible to impose an ideal semantic topology on the network. All relevant peers and thus all bibliographic descriptions can be found in a deterministic manner, as the query is simply routed along the route which corresponds to the shortest path in the ACM topic hierarchy. At each hop the query is forwarded to exactly one peer until the relevant peer is reached. The number of messages required per query is therefore the length of the shortest path from the topic of expertise of the originating peer to that of the topic of the query subject. The precision of the peer selection increases to the maximum when arriving at the eight hop, which is the maximum possible length of a shortest path in the ACM topic hierarchy. Accordingly, the maximum number of messages (Figure 4) required is also eight.

6 Related Work

The idea of expertise based matching for peer selection using ontologies is similar to that of capability based matching as described in [16], where specifications of requests are matched against a set of capabilities of agents or services. Capability based matching has recently also been applied for matching of Web Services, e.g. [9].

Another approach, which does a semantic comparison between a query and a peer's context comes from [4]. They propose a Peer-to-Peer architecture, implemented as their 'KEx' system where queries can be accompanied with a 'focus' which is a part of an ontology, e.g. a small taxonomy. When a peer receives a query, its matching algorithm tries to match the focus of the query semantically and syntactically. The syntactic matching process is straight-forward by using an indexer to search for the occurrence of specific keywords into the set of documents owned by the provider. For the semantic matching a context matching algorithm is used that tries to find a correlation between a provider's context and the query focus. In particular the matching algorithm tries to find the focus in the provider's context that has a relevant semantic relation with the one sent by the seeker. Related documents that fit the focus are returned as results. If the focus points to other peers, the provider will propagate the query. The big strength of this approach is that it does not make the assumption that the ontologies should be equal and shared by all the peers, contrary to our approach. The advantage of our approach however is that it is much easier to calculate the similarity between a query's subject and the expertise of a peer.

pSearch [17] distributes document indices through the P2P network based on document semantics generated by Latent Semantic Indexing (LSI) [3]. LSI represents documents and queries as vectors in a Cartesian space and measures the similarity between a query and a document as the cosine of the angle between their vector representations. pSearch is organized as a Content-addressable network (CAN) [19]. CANs provide a distributed hash table (DHT) abstraction also distributed over a Cartesian space. The combination of the LSI representation and their network organization, the search cost (in terms of different nodes searched and data transmitted) for a given query is reduced, since the indices of semantically related documents are likely to be co-located in the network. Although the pSearch approach seems to work very well for finding documents close to a query, the vector dimensionality and the corresponding concepts for each place in the vectors need to be known beforehand. In their experiments they used a vector with a dimensionality around a few hundred concepts. This means that all the documents in the system can only be identified and matched on these corresponding concepts. In other words, the network topology is directly connected and therefore limited by the number of concepts. This is contrary to our approach where we don't make any assumption about the network topology.

A completely different approach for finding experts in a network comes from social network analysis. ReferralWeb [8] uses the social network to make a search more focused and effective. ReferralWeb attempts to uncover the existing social networks by data mining public documents found on the WWW. Such sources can include links found on home pages, lists of co-authors in technical papers and citations of papers, exchanges between individuals recorded in news archives, and organization charts. Their simulation experiments showed that automatically generated referrals can be highly successful in

locating experts in a large network. Experiments performed by [19] show that when referrals are considered, better answers are found in terms of precision. They also show that it is possible to let the system evolve to a situation where peers with similar expertise and interest are grouped close towards each other, according to their own similarity function. It is probable that the number of messages needed for getting an answer on a query decreases when the system evolves, but unfortunately that isn't shown by their experiments. The main difference with our approach is that their peers express queries and expertise in a vector, in which the similarity is based on taking the cosine of both vectors.

[12] presents schema-based Peer-to-Peer networks and the use of super-peer based topologies for these networks, in which peers are organized in hypercubes. [11] shows how this schema-based approach can be used to create Semantic Overlay Clusters in a scientific Peer-to-Peer network with a small set of metadata attributes that describe the documents in the network. In contrast, the approach in our system is completely decentralized in the sense that it does not rely on super-peers.

7 Conclusions and Future Work

Summary: In this paper we have presented a model for expertise-based peer selection, in which a semantic topology among the peers is created by advertising the expertise of the peers. We have shown how the model can be applied in a bibliographic scenario. Simulation experiments that we performed with this bibliographic scenario show the following results:

- Using expertise-based peer selection can increase the performance of the peer selection by an order of magnitude (result R1).
- However, if expertise-based peer selection uses simple exact matching, the recall drops to unacceptable levels. It is necessary to use an ontology-based similarity measure as the basis for expertise-based matching (result R2).
- An advertising strategy where peers only accept advertisements that are semantically close to their own profile (i.e. that are in their semantic neighborhood) is a simple and effective way of creating a semantic topology. This semantic topology allows to forward queries along the gradient of increasing semantic similarity (result R3).
- The above results depend on how closely the semantic topology of the network mirrors the structure of the ontology. All relevant performance measure reach their optimal value when the network is organised exactly according to the structure of the topology (result R4). Although this situation is idealised and in will in practice not be achievable, the experiment serves to confirm our intuitions on this.

Summarizing, in simulation experiments we have shown that expertise-based peer selection combined with ontology-based matching outperforms both random peer selection and selection based on exact matches, and that this performance increase grows when the semantic topologies more closely mirrors the domain ontology.

Limiting assumptions: We have made a number of simplifying assumptions in our experiments. We review these assumptions, and the likely impact their relaxation may have on our results:

- **A single ontology:** clearly, the assumption that all peers agree on the use of single ontology is not in all cases realistic. We already have work in progress which allows us to relax this constraint. We expect that differences in ontologies used by different peers will *lower* our results, since the computation of the semantic distance between peers becomes less reliable across different ontologies.
- **A static semantic topology:** in our experiments, the semantic topology is determined once, during an initial advertising round, and is not adapted any further during the lifetime of the experiment. The work in [18] shows how the topology can be adjusted based on the exchange of queries and answers. We expect that such a self-adjusting network will *improve* our results, since the semantic topology will converge better towards the structure of the underlying ontology than our current one-shot advertising allows.
- **Static content distribution:** in our experiments, content was assigned statically to peers, while in a realistic network, the content of different peers is likely to evolve over the lifetime of the network. Since such changing content will also induce changes in the expertise profile of the peers, we expect that this assumption can only be relaxed in the presence of self-adjusting semantic topologies (as mentioned in the previous point). Again, we have work in progress to relax this assumption.

Future work: Besides relaxing the above assumptions, there are many other fruitful directions in which this work can be taken:

- **More complex expertise models.** The expertise model presented for the bibliographic scenario is a fairly simple one, based on the ACM topic hierarchy. Other domains may require more complex expertise models with different similarity functions. One option would be, for example, to extend the expertise model with quantitative measures to indicate how much information for a certain topic of expertise is available on the peer.
- **Merge semantic and network topology.**
  So far we have considered the semantic topology to be independent of the underlying network topology. It would however be interesting to use, for example, the extensibility mechanisms of the JXTA platform to extend its default mechanisms for discovery and query routing with the methods presented in this paper.
- **Field Experiment.** To verify the results of the simulation experiments in the real world, the model proposed in this paper is currently implemented in the Bibster system[4] and evaluated in the bibliographic scenario with a field experiment [5].

Acknowledgments. Research reported in this paper has been partially financed by the EU in the IST project SWAP (IST-2001-34103). We would like to thank our colleagues for fruitful discussions.

[4] http://bibster.semanticweb.org

References

1. Benjamin Ahlborn, Wolfgang Nejdl, and Wolf Siberski. OAI-P2P: A peer-to-peer network for open archives. In *2002 International Conference on Parallel Processing Workshops (ICPPW'02)*, 2002.
2. Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 2001.
3. Michael W. Berry, Zlatko Drmac, and Elizabeth R. Jessup. Matrices, vector spaces, and information retrieval. In *SIAM Review*, pages 335–362, 1999.
4. Matteo Bonifacio, Roberta Cuel, Gianluca Mameli, and Michele Nori. A peer-to-peer architecture for distributed knowledge management. In *Proceedings of the 3rd International Symposium on Multi-Agent Systems, Large Complex Systems, and E-Businesses MALCEB'2002*, 2002.
5. Jeen Broekstra, Marc Ehrig, Peter Haase, Frank van Harmelen, Maarten Menken, Peter Mika, Björn Schnizler, and Ronny Siebes. Bibster - a semantics-based bibliographic peer-to-peer system. In *Proceedings of the WWW'04 Workshop on Semantics in Peer-to-Peer and Grid Computing, New York, 2004*, 2004.
6. Jeen Broekstra and Arjohn Kampman. SeRQL: An RDF Query and Transformation Language. Submitted to the International Semantic Web Conference, ISWC 2004, 2004. See also http://www.openrdf.org/doc/SeRQLmanual.html.
7. Marc Ehrig, Christoph Schmitz, Steffen Staab, Julien Tane, and Christoph Tempich. Towards evaluation of peer-to-peer-based distributed knowledge management systems. In *Proceedings of the AAAI Spring Symposium "Agent-Mediated Knowledge Management (AMKM-2003)"*, 2003.
8. Bart Selman Henry Kautz and Mehul Shah. Referralweb: Combining social networks and collaborative filtering. In *Communications of the ACM*, March 1997.
9. Lei Li and Ian Horrocks. A software framework for matchmaking based on semantic web technology. In *Proceedings of the Twelfth International World Wide Web Conference (WWW 2003)*, pages 331–339. ACM, 2003.
10. Yuhua Li, Zuhair A. Bandar, and David McLean. An approach for measuring semantic similarity between words using multiple information sources. *Transactions on Knowledge and Data Engineering*, 15(4):871–882, July/August 2003.
11. Alexander Löser, Martin Wolpers, Wolf Siberski, and Wolfgang Nejdl. Efficient data store discovery in a scientific P2P network. In N. Ashish and C. Goble, editors, *Proceedings of the WS on Semantic Web Technologies for Searching and Retrieving Scientific Data*, CEUR WS 83, 2003. Colocated with the 2. ISWC-03.
12. Wolfgang Nejdl, Martin Wolpers, Wolf Siberski, Christoph Schmitz, Mario Schlosser, Ingo Brunkhorst, and Alexander Löser. Super-peer-based routing and clustering strategies for rdf-based peer-to-peer networks. In *Proceedings of the 12th International World Wide Web Conference, Budapest, Hungary, May 2003.*, 2003.
13. Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.
14. Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. In *Proceedings of ACM SIGCOMM '01*, 2001.
15. Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for Internet applications. In *Proceedings of the ACM SIGCOMM '01*, 2001.
16. Katia Sycara, Keith Decker, and Mike Williamson. Middle-agents for the internet. In *Proceedings of IJCAI-97*, January 1997.

17. Chunqiang Tang, Zhichen Xu, and Sandhya Dwarkadas. Peer-to-peer information retrieval using self-organizing semantic overlay networks. In *Proceedings of the ACM SIGCOMM Conference*, Karlsruhe, Germany, August 2003.
18. Christoph Tempich, Steffen Staab, and A. Wranik. REMINDIN': Semantic query routing in peer-to-peer networks based on social metaphors. In *Proceedings of the 13th Int. World Wide Web Conference, WWW 2004*, 2004.
19. Pinar Yolum and Munindar P. Singh. Dynamic communities in referral networks. *Web Intelligence and Agent Systems*, 1(2):105–116, 2003.