Bibster – A Semantics-Based Bibliographic Peer-to-Peer System

Peter Haase[1], Jeen Broekstra[3], Marc Ehrig[1], Maarten Menken[2], Peter Mika[2], Mariusz Olko[4], Michal Plechawski[4], Pawel Pyszlak[4], Björn Schnizler[1], Ronny Siebes[2], Steffen Staab[1], and Christoph Tempich[1]

[1] Institute AIFB, University of Karlsruhe, D-76128 Karlsruhe, Germany
{ehrig, haase, staab, tempich}@aifb.uni-karlsruhe.de, schnizler@iw.uka.de
[2] Vrije Universiteit Amsterdam, The Netherlands
{mrmenken, pmika, ronny}@cs.vu.nl
[3] Aduna, Amersfoort, The Netherlands jeen@aduna.biz
[4] Empolis, Warsaw, Poland {pap, mpl}@empolis.pl

Abstract. This paper describes the design and implementation of Bibster, a Peer-to-Peer system for exchanging bibliographic data among researchers. Bibster exploits ontologies in data storage, query formulation, query routing and answer presentation: When bibliographic entries are made available for use in Bibster, they are structured and classified according to two different ontologies. This ontological structure is then exploited to help users formulate their queries. Subsequently, the ontologies are used to improve query routing across the Peer-to-Peer network. Finally, the ontologies are used to post-process the returned answers in order to do duplicate detection. The paper describes each of these ontology-based aspects of Bibster. Bibster is a fully implemented open source solution built on top of the JXTA platform.

1 Introduction

The advantages of Peer-to-Peer architectures over centralized approaches have been well advertised, and to some extent realized in existing applications: no centralized server (thus avoiding a bottleneck for both computational performance and information update), robustness against failure of any single component, scalability both in data volumes and the number of connected parties. However, besides being the solution to many problems, the large degree of distribution of Peer-to-Peer systems is also the cause of a number of new problems: The lack of a single coherent schema for organizing information sources across the Peer-to-Peer network hampers the formulation of search queries, duplication of information across the network results in many duplicate answers to a single query, and answers to a single query often require the integration of information residing at different, independent and uncoordinated peers.Finally, query routing and network topology (which peers to connect to, and which peers to send/forward queries to) are significant problems.

S.A. McIlraith et al. (Eds.): ISWC 2004, LNCS 3298, pp. 122–136, 2004.

The research community has recently turned to the use of semantics in Peer-to-Peer networks to alleviate these problems [1], [2], [3]. The use of semantic descriptions of datasources stored by peers and indeed of semantic descriptions of peers themselves helps in formulating queries such that they can be understood by other peers, in merging the answers received from other peers, and in routing queries across the network. In particular, the use of ontologies and of Semantic Web technologies has been identified as promising for Peer-to-Peer systems.

The scenario that we have envisioned is that researchers share bibliographic metadata in a community with a Peer-to-Peer system. The data may have been obtained from local BibTeX files or from bibliography servers like the DBLP database[1] or CiteSeer[2]. As one may easily recognize, this scenario (like some others that we do not elaborate upon in this paper) exhibits two characteristics that strongly require a semantics-based Peer-to-Peer system.

First, a centralized solution does not exist and cannot exist, because of the multitude of informal workshops that researchers refer to, but that do not show up in centralized resources such as DBLP. Any such centralized resource will only cover a limited scientific community. For example, DBLP covers a lot of Artificial Intelligence, but almost no Knowledge Management, whereas a lot of work is being done in the overlap of these two fields. At the same time, many individual researchers are willing to share their resources, provided they do not have to invest work in doing so.

Second, the use of Semantic Web technology is crucial in this setting. Although a small common-core ontology of bibliographic information exists (title, author/editor, etc), much of this information is very volatile and users define arbitrary add-ons, for example to include URLs or abstracts of publications.

In this paper we will describe the design of the Bibster system and emphasize the semantic components and their use: semantic extraction of bibliographic metadata in section 4, semantic querying in section 5, peer selection using semantic topologies in section 6, and semantic duplicate detection in section 7. Furthermore, Bibster has been fully implemented and we present evaluation results from a field experiment in section 8.

2 Major Bibster Use Cases

Bibster is aimed at researchers that share bibliographic metadata. Requirements for Bibster must include capabilities that support their daily work. Researchers may want to:

1. query a single specific peer (e.g. their own computer, because it is sometimes hard to find the right entry there), a specific set of peers (e.g. all colleagues at an institute) or the entire network of peers (to obtain the maximal recall at the price of low precision).

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

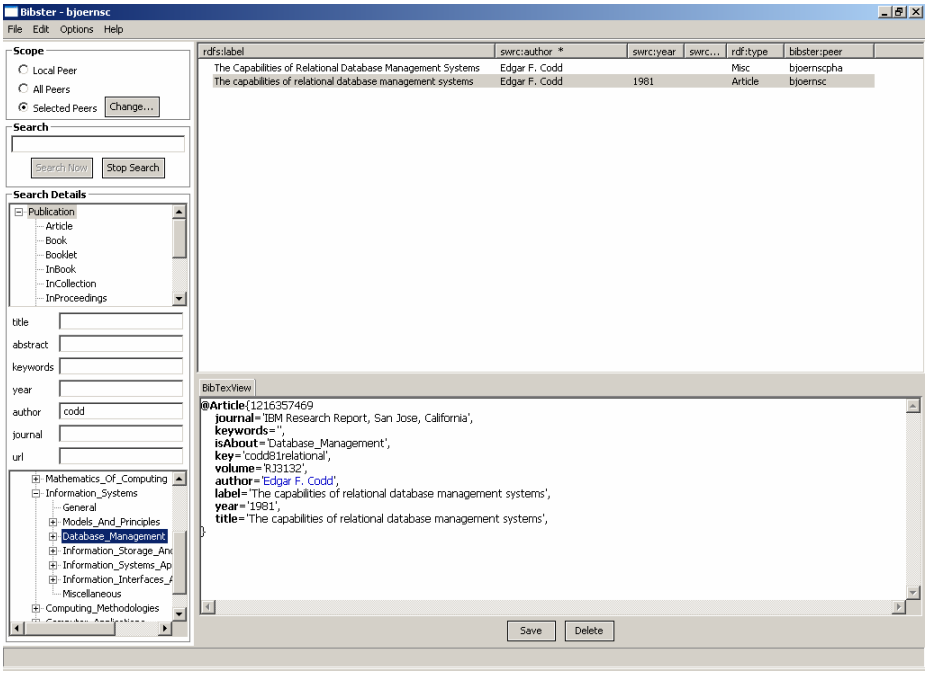

Fig. 1. Searching for publications about database management authored by Codd

2. search for bibliographic entries using simple keyword searches, but also more advanced, semantic searches, e.g. for publications of a special type, with specific attribute values, or about a certain topic.
3. integrate results of a query into a local knowledge base for future use. Such data may in turn be used to answer queries by other peers. They may also be interested in updating items that are already locally stored with additional information about these items obtained from other peers.

The screenshot in figure 1 partially indicates how these use cases are realized in Bibster. The *Scope* widget allows for defining the targeted peers, the *Search* and *Search Details* widgets allow for keyword and semantic search; *Results Table* and *BibtexView* widgets allow for browsing and re-using query results. The query results are visualized in a list grouped by duplicates. They may be integrated into the local repository or exported in formats such as BibTeX and HTML.

3 Design of Bibster

3.1 Ontologies in Bibster

Ontologies are crucial throughout the usage of Bibster, viz. for importing data, formulating queries, routing queries, and processing answers.

Firstly, the system enables users to import their own bibliographic metadata into a local repository. Bibliographic entries made available to Bibster by a user (cf. section 4) are automatically aligned to two common ontologies: The first ontology (SWRC[3]) describes different generic aspects of bibliographic metadata (and would be valid across many different research domains), the second ontology (ACM Topic Hierarchy[4]) describes specific categories of literature for the Computer Science domain.

Secondly, queries are formulated in terms of the two ontologies: Queries may concern fields like author, publication type, etc. (using terms from the SWRC ontology) or queries may concern specific Computer Science terms (using the ACM Topic Hierarchy).

Thirdly, queries are routed through the network depending on the expertise models of the peers describing which concepts from the ACM ontology a peer can answer queries on. A matching function determines how closely the semantic content of a query matches the expertise model of a peer. Routing is then done on the basis of this semantic ranking.

Finally, answers are returned for a query. Due to the distributed nature and potentially large size of the Peer-to-Peer network, this answer set might be very large, and contain many duplicate answers. Because of the semistructured nature of bibliographic metadata, such duplicates are often not exactly identical copies. Ontologies help to measure the semantic similarity between the different answers and to remove apparent duplicates as identified by the similarity function.

3.2 Bibster Architecture and Modules

The Bibster system has been implemented as an instance of the SWAP System architecture as introduced in [1]. Figure 2 shows a high-level design of the architecture of a single node in the Peer-to-Peer system. We will now briefly present the individual components as instantiated for the Bibster system.

Communication Adapter: This component is responsible for the network communication between peers. It serves as a transport layer for other parts of the system, for sending and forwarding queries. It hides and encapsulates all low-level communication details from the rest of the system. In the specific implementation of the Bibster system we use JXTA as the communication platform.

Knowledge Sources: The knowledge sources in the Bibster system are sources of bibliographic metadata, such as BibTeX files stored locally in the file system of the user.

Knowledge Source Integrator: The Knowledge Source Integrator is responsible for the extraction and integration of internal and external knowledge sources into the Local Node Repository. In section 4 we describe the process of semantic extraction from BibTeX files. In section 7 we explain how the knowledge of local and remote sources can be merged, i.e. how duplicate query results are detected.

[3] http://www.semanticweb.org/ontologies/swrc-onto-2001-12-11.daml
[4] http://www.acm.org/class/1998/

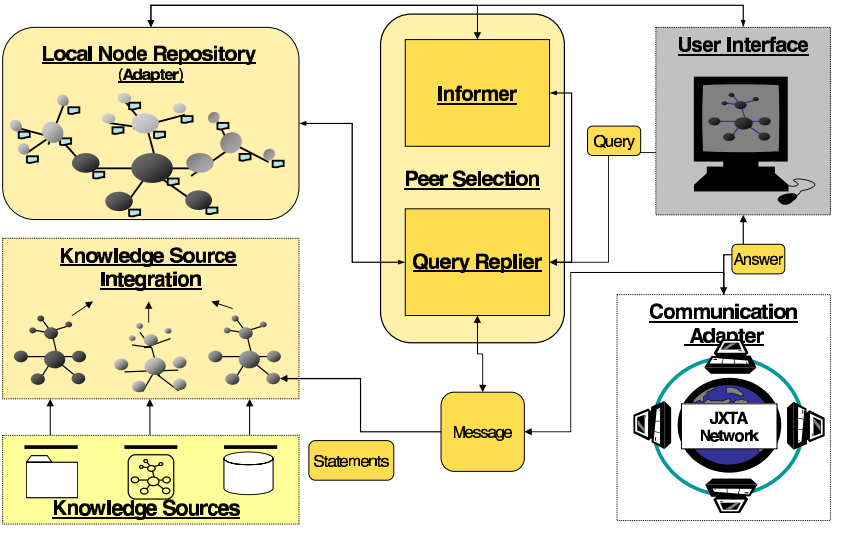

Fig. 2. SWAP System Architecture

Local Node Repository: In order to manage its information models and views as well as information acquired from the network, each peer maintains a Local Node Repository providing the following functionality: (1) Mediate between views and stored information, (2) support query formulation and processing, (3) specify the peer's interface to the network, and (4) provide the basis for peer ranking and selection. In the Bibster system, the Local Node Repository is based on the RDF-S Repository Sesame [4]. The query language SeRQL is used to formulate semantic queries against the Local Node Repository, as described in section 5.

Informer: The task of the Informer is to proactively advertise the available knowledge of a peer in the Peer-to-Peer network and to discover peers with knowledge that may be relevant for answering the user's queries. This is realized by sending advertisements about the expertise of a peer. In the Bibster system, these expertise descriptions contain a set of topics that the peer is an expert in. Peers may accept – i.e. remember – these advertisements, thus creating a semantic link to the other peer. These semantic links form a semantic topology, which is the basis for intelligent query routing (cf. section 6 for details).

Query Replier: The Query Replier is the coordinating component controlling the process of distributing queries. It receives queries from the User Interface or from other peers. Either way it tries to answer the query or distribute it further according to the content of the query. The decision to which peers a query should be sent is based on the knowledge about the expertise of other peers.

User Interface: The User Interface (Figure 1) allows the user to import, create and edit bibliographic metadata as well as to easily formulate queries.

4 Semantic Extraction of Bibliographic Metadata

Many researchers have accumulated extensive collections of BibTeX files for their bibliographic references. However, these files are semi-structured and thus single attributes may be missing or may not be interpreted correctly. For interchanging bibliographic data in a semantically based Peer-to-Peer network it has to be represented in a structured and formal way. *BibToOnto* is a component of Bibster for extracting explicit knowledge of bibliographic items. Plain BibTeX files are transformed into an ontology based knowledge representation.

The target ontology is the Semantic Web Research Community Ontology (SWRC), which models among others a research community, its researchers, topics, publications, tools, and properties between them. The SWRC ontology defines a shared and common domain theory which helps users and machines to communicate concisely and supports the exchange of semantics.

BibToOnto automatically classifies bibliographic entries according to the ACM topic hierarchy. Additionally, it is possible to reclassify the entries manually in the user interface of Bibster. The ACM topic hierarchy is a standard schema for describing and categorizing computer science literature. It covers 1287 topics of the computer science domain. In addition to the sub- and supertopic relations, it also provides information about related topics.

The following example shows a transformation of a BibTeX entry to a SWRC ontology based item. The result[5] is represented as an RDF graph in figure 3.

Example 1.
```
@ARTICLE{codd70relational
        author   = "Edgar F. Codd",
        year     = "1970",
        title    = "A relational model for large shared data banks",
        journal  = "Communications of ACM",
        volume   = "13",
        number   = "6",
        pages    = "377--387"}
```

5 Semantic Querying

Each peer node in the Bibster system manages a local RDF repository with bibliographic data extracted by BibToOnto or integrated from other peers. The query language interface to the local RDF repository is SeRQL [5].

SeRQL (Sesame RDF Query Language) is an RDF/RDF-S query language that was developed in the context of the SWAP project to address practical requirements that were not sufficiently met by other query languages.

SeRQL Design Principles. Several characteristics are urgently required in the context of Bibster (but also of many other systems) for an RDF/RDF-S query language like SeRQL. In particular, it must:

[5] For better readability we used a concatenation of the author name and the title of the publication as a URI in this example. In the Bibster system however we calculate hash codes over all attribute values to guarantee the uniqueness of URIs.

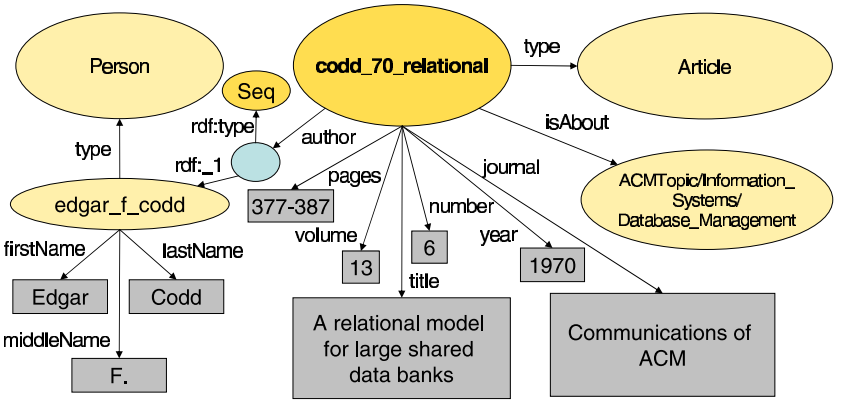

Fig. 3. SWRC Sample Metadata

1. be functional such that each query returns a RDF graph, which may be integrated into the local repository or queried again,
2. be aware of the (optional) schema,
3. let the user formulate path expressions for navigating the RDF graph, e.g. the combination of SWRC and ACM topic hierarchy,
4. be able to deal with *optional* values, e.g. a publisher field may be given or not.

Without showing all capabilities of SeRQL in full detail[6], we briefly show how SeRQL queries are composed, and how tasks in the Bibster system are performed using SeRQL.

(1) SeRQL uses a `select-from-where` or `construct-from-where` filter, where the `select` or `construct` clauses specify projections, the `from` clause specifies a graph match template (by means of path expressions), and the `where` clause allows the definition of additional boolean constraints on matched values in the path expressions.

(2) SeRQL takes the RDF schema into account by mapping from the given graph to its formal model.

(3) When navigating the RDF graph, SeRQL exploits the formal semantics of path labels. For example, `<rdfs:subClassOf>` is interpreted as a reflexive transitive relation and upward inheritance of instances is interpreted (through the `<rdf:type>` relation (cf. [5] for full details).

(4) Bibtex entries may be incomplete. SeRQL allows to distinguish between optional and required elements in the query and, hence, is flexible enough to deal with these circumstances.

A Querying Scenario. In our running example, a researcher is querying for journal articles written by the author Codd about database management. Internally, this request is formulated as a SeRQL query that looks as follows:

[6] See `http://www.openrdf.org/doc/SeRQLmanual.html` for a complete overview.

Example 2.

```
construct distinct
    {s} prop {val};
        <rdf:type> {t};
        <swrc:author> {x} <rdf:type> {<rdf:Seq>};
                          <rdfs:member> {author} prop_author {val_author}
from
    {s} <serql:directType> {t};
        <rdf:type> {<swrc:Article>};
        prop {val};
        <swrc:isAbout> {<acm:ACMTopic/Information_Systems/Database_Management>};
        <swrc:author> {x} <rdfs:member> {} <swrc:lastName> {lname},
        [{x} <rdfs:member> {author} prop_author {val_author} ]
where prop != <rdf:type> and lname like "Codd"
using namespace
    swrc = <!http://www.semanticweb.org/ontologies/swrc-onto-2001-12-11.daml#>,
    acm =  <!http://daml.umbc.edu/ontologies/classification#>
```

Compare the structure of the from-clause to the representation of the RDF graph given in figure 3. The from-clause retrieves not only the identifier for the particular journal entry ("codd_70_relational", matched by `s`), but also the graph structure surrounding it, which essentially gives the entry its meaning: the name of the author, the type of publication, the year it was published, the number of pages, etc. Also, if the first and middle names of an author are known, the query retrieves those (but it does not fail if these are not known).

The use of schema-awareness is evident in the use of typing information on `s`: `s` need not only be of type `swrc:Article`, we also retrieve its *specific* (or *direct*) type. Being functional plays a role as well: A graph transformation is used to create a query result that can be easily processed to be given back to the user through the GUI.

6 Expertise Based Peer Selection

The scalability of a Peer-to-Peer network is essentially determined by the way how queries are propagated in the network. Peer-to-Peer networks that broadcast all queries to all peers do not scale – intelligent query routing and network topologies are required to be able to route queries to a relevant subset of peers that are able to answer the queries.

Modern routing protocols like Chord [6] and CAN [7] allow for sophisticated query routing based on distributed indices. More recently, in the Semantic Web context, schema based Peer-to-Peer networks such as the one described in [8] have emerged based on complex, extensible semantic descriptions of resources. They allow for complex queries against these metadata instead of simple keyword-based queries. Another semantic-based approach is pSearch [9], a decentralized non-flooding Peer-to-Peer information retrieval system. pSearch distributes document indices through the Peer-to-Peer network based on document semantics generated by Latent Semantic Indexing (LSI). The search cost (in terms of nodes searched and data transmitted) for a given query is thereby reduced, since the indices of semantically related documents are likely to be co-located in the network. Here we give an overview of the model of expertise based peer selection

as proposed in [10] and how it is used in the Bibster system. In this model, peers use a shared ontology to advertise semantic descriptions of their expertise in the Peer-to-Peer network. The knowledge about the expertise of other peers forms a 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 the ones that have a good chance of answering it.

Semantic Description of Expertise

Peers. The Peer-to-Peer network consists of a set of peers P. Every peer $p \in P$ has a Local Node Repository, which stores the bibliographic metadata.

Common Ontology. The peers share an ontology O, which is used for describing the expertise of peers and the subject of queries. In our case, O is the ACM topic hierarchy that contains a set of topics T.

Expertise. An expertise description is an abstract, semantic description of the Local Node Repository of a peer based on the shared ontology O. The expertise E of a peer is thus 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 $A \subseteq P \times E$ are used to promote descriptions of the expertise of peers in the network. An advertisement $a \in A$ associates a peer p with an expertise e. Peers decide autonomously, without central control, whom to promote advertisements to and which advertisements to accept. This decision is based on the semantic similarity between expertise descriptions.

Matching and Peer Selection

Queries. Queries $q \in Q$ are posed by a user and are evaluated against the local node repositories of the peers. First a peer evaluates the query against its local node repository and then decides which peers the query should be forwarded to.

Subjects. A subject is an abstraction of a given query q expressed in terms of the common ontology. The subject specifies the required expertise to answer the query. In our scenario, the subjects of queries are defined as $S \subseteq 2^T$, each s is the set of ACM topics that are referenced in the query. E.g., the extracted subject of the query in example 2 would be *Information Systems/Database Management*.

Similarity Function. The similarity function $Sim : S \times E \mapsto [0, 1]$ 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. Sim is used for determining to which peers a query should be forwarded. In Bibster, the similarity function Sim_{Topics} 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 the 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. [11] have compared different similarity measures and have shown that for measuring the similarity between concepts in a hierarchical structured semantic network, like the ACM topic hierarchy, the following similarity measure yields the best results:

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

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: $\alpha = 0.2$, $\beta = 0.6$.

Peer Selection Algorithm. The peer selection algorithm returns a ranked set of peers, where the rank value is equal to the similarity value provided by the similarity function. Therefore, peers that have an expertise more similar to that of the subject of the query will have a higher rank. 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. In the Bibster system we select the best n peers that have not yet received the query along the message path, where n can be specified. A maximum number of hops (set to 4 for the field experiment) further limits the forwarding of queries.

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 do not 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.

7 Semantic Duplicate Detection

When querying the Bibster network one receives a large number of results with an often high number of duplicates. This is due to the fact that we do not

have a centralized but many distributed local repositories. Furthermore, the representation of the metadata is very heterogeneous and possibly even contradicting. To enable an efficient and easily usable system Bibster presents query results grouping duplicates together. Duplicates in Bibster are bibliographic entries which refer to the same publication, person, or organization in the real world, but are modelled as different resources. Bibster uses specific similarity functions to recognize two resources as being duplicates.

Similarity Function. A similarity function for RDF resources R of the local node repository is a function $sim : R \times R \rightarrow [0..1]$.

For each resource type (publication, person, organization), we have compiled a set of specific features used to assess the similarity between two of its instances. For instance, publications are assessed based on their titles, publication types, authors, years, ACM topics, etc. For each of the features we use different *individual similarity functions*, which are grouped as follows:

The *data value level* focuses on comparisons of data values, which in RDF are represented as typed literals. For example, to determine the similarity of data values of type string (e.g. to compare the last names of persons) we use the *syntactic similarity* of [12]. At the *graph structure level* we check how resources are related to each other. For example, a publication resource is linked to person resources, e.g. authors. Thus we can compare two publications on the basis of the similarity of the sets of authors. This feature alone is not deciding, but it supports the hypothesis of having duplicate entries.

Similarity measures at the *ontology level* extend the ones at the graph structure level by ontology specific characteristics. To determine the similarity of two publications based on their topics we make use of the ACM topic hierarchy. We apply the hierarchical similarity function as presented earlier in this paper [13].

Applying background *knowledge about a specific domain*, we can define more appropriate similarity functions. For example, in the SWRC domain ontology there are many subconcepts of publications: articles, books, and technical reports to just name a few. Unknown publication types are often provided as Misc. We can thus define a function that returns a value of 1 if the publication type is identical, a value of e.g. 0.75 if Misc is one of it, and 0 otherwise.

From the variety of individual similarity functions, an overall value is obtained with an aggregated similarity function, using a weighted average over the individual functions. For Bibster, the weights have been assigned based on experiments with sample data. More precisely, several duplicates were detected manually. From these training duplicates the weights were adjusted to achieve a maximal f-measure (combination of precision and recall) value.

The Duplicate Relation. As duplicates we consider those pairs of resources whose similarity is larger than a certain threshold

$$t \in [0..1] : D_t := \{(x, y) | sim(x, y) \geq t\}$$

If we assume that the duplicate relation is transitive, we can define the transitive closure as:

$$TC(D_t) := \{(x, z) | (x, y) \in D_t \wedge (y, z) \in D_t\}$$

This transitive closure essentially represents clusters of semantically similar resources.

Resource Merging. Instead of presenting all individual resources of the query result, duplicates are visualized as one, merged, resource. The merged resources comprise the union of statements of the individuals identified as duplicates. In the case of conflicting property values, we apply heuristics for the merging of resources (e.g. for booktitles to select the most detailed value with the least abbreviations).

8 Results

The Bibster system that implements the methods presented in this paper has been evaluated by means of a public field experiment. The user actions and system events are continuously logged and analyzed to evaluate the user behavior and system performance. We have analyzed the results for a period of two months (June and July 2004) and have obtained the following interesting results: A total of 146 peers from various organizations spread mainly over Europe and North America used the Bibster system. The users shared more than 70000 bibliographic entries. While seventeen peers shared more than 1000 items each, accounting for 84% of the total content, a lot of peers provided only little content or were "free-riding".

The users performed a total of 1782 queries. The SWRC ontology was used for about half of all queries, mainly for the purpose to search for special types of publications (e.g. only for articles) or for publications of a given author. In 348 queries the users asked for topics of the ACM topic hierarchy. Thereby it is obvious that the users are accepting the ontology based searching capabilities and that there is a benefit for them in using these ontologies.

With respect to query routing, with the expertise based peer selection we were able to reduce the number of query messages by about 50 percent, while retaining the same recall of documents compared with a naive broadcasting approach. Figure 4 shows the number of forwarded query messages sent per query and the precision of the peer selection (the percentage of the reached peers that actually provided answers to a given query). Although we have shown an improvement in the performance, the results also show that with a network of the size as in the field experiment, a naive approach is also acceptable. On the other hand, with a growing number of peers, query routing and peer selection becomes critical: In simulation experiments with larger peer networks with thousands of peers, we have shown improvements in the order of one magnitude in terms of recall of documents and relevant peers [10].

Lessons Learned

This section summarizes some experiences we have gained from the development and application of Bibster.

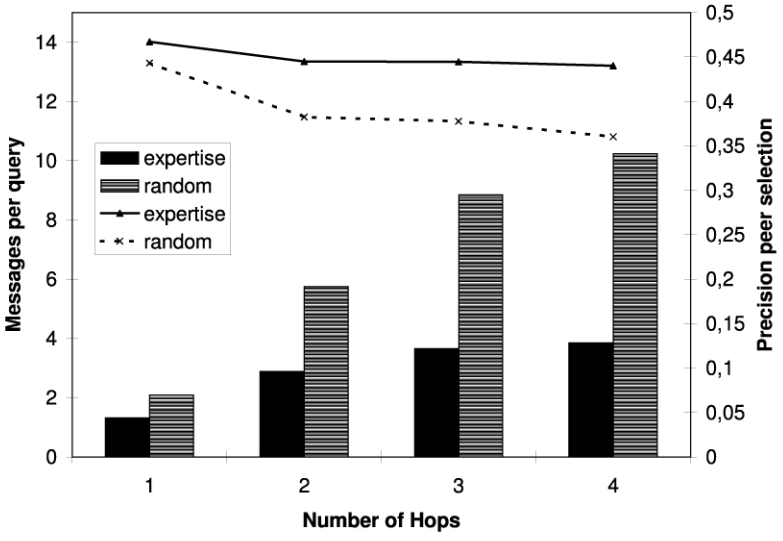

Fig. 4. Number of Messages and Precision of Peer Selection

For Bibster and similar applications the usage of Semantic Web technologies and ontologies provide an *added value* — in fact it is almost a strict requirement given its semi-structured, volatile data structures. Semantic structures serve important user concerns like high quality duplicate detection or comprehensive searching capabilities.

Unsurprisingly, in small networks with *small user groups*, intelligent query routing is not a major issue. While it is beneficial to direct queries to specific peers known to the user (a trust issue!), advanced routing algorithms may only be beneficial for a much larger number of users in a network. Based on our experience we now conjecture that content-based routing and trust issues will have to converge for such larger networks, too.

9 Related Work

In the previous sections, related work on the individual aspects of semantics-based Peer-to-Peer technology has already been discussed. Therefore in this section our study of related work focuses on complete systems. Edutella (*cf. eg.* [8]) is a Peer-to-Peer system based on the JXTA platform, which offers similar base functionality as the SWAP system. The Edutella network focuses on the exchange of learning material. They use super-peer based topologies, in which peers are organized in hypercubes to route queries. In contrast to their work, Bibster is embedded in the general SWAP architecture and a running application.

[14] describes the design of a Peer-to-Peer network for open archives, where data providers, i.e. research institutes, form a Peer-to-Peer network which sup-

ports distributed search over all the connected metadata repositories. This scenario which is similar to our bibliographic Peer-to-Peer scenario, however, their system has not been implemented up to this point.

P-Grid [15] is a structured, yet fully-decentralized Peer-to-Peer system based on a virtual distributed search tree. It aims at providing load-balancing and fault-tolerance, assuming that peers fail frequently and are online with low probability. P-Grid also considers updates with an update algorithm based rumor spreading.

The DFN Science-to-Science (S2S) [16] system enhances content based searching by using peer-to-peer technology to make locally generated indexes accessible in an ad hoc manner. Whereas Bibster is fully distributed, S2S uses a kind of super peers (Search Hubs) to route queries and cache information.

Various systems address the issue of heterogeneity in Peer-to-Peer systems on the schema level, such as the Piazza peer data management system [17], which allows for information sharing with different schemas relying on local mappings between schemas.

10 Conclusion

In this paper, we have described the design and implementation of Bibster, a semantics-based Peer-to-Peer system for the exchange of bibliographic metadata between researchers. For this purpose, Bibster exploits lightweight ontologies, expressed in RDF Schema in all its crucial aspects: data-organisation, query formulation, query routing and duplicate detection. To our knowledge, Bibster now constitutes the first ontology-based Peer-to-Peer systems ready for fielded deployment.

In general, there are interesting alternatives for each of the different aspects (e.g., [2] for querying or [15] for query routing) and, actually, we are still exploring multiple approaches to optimize the overall system (e.g., [18] for query routing). In practice, however, it constitutes a major challenge to integrate these different components into a coherent system like Bibster.

The next steps in the development of Bibster are, *(i)* its optimization (e.g., manual query optimization), *(ii)* its spreading to further user groups and, *(iii)* the extension of Bibster to better account for personalized semantic structures, based on the two common core ontologies, e.g. peer-local extensions of the ACM topic hierarchy.

The reader may find further reading material on Bibster, its underlying technologies and related material in the open available project deliverable documentation at `http://swap.semanticweb.org/` and `http://bibster.semanticweb.org/`.

Acknowledgments. Research reported in this paper has been partially financed by the EU in the IST projects SWAP (IST-2001-34103) and SEKT (IST-2003-506826). We would like to thank our colleagues for fruitful discussions.

References

1. Broekstra, J., Ehrig, M., Haase, P., van Harmelen, F., Kampman, A., Sabou, M., Siebes, R., Staab, S., Stuckenschmidt, H., Tempich, C.: A metadata model for semantics-based peer-to-peer systems. In: Proceedings of the WWW'03 Workshop on Semantics in Peer-to-Peer and Grid Computing. (2003)
2. Nejdl, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmér, M., Risch, T.: Edutella: A P2P networking infrastructure based on rdf. In: Proceedings to the Eleventh International World Wide Web Conference. (2002)
3. Castano, A., Ferrara, S., Montanelli, S., Pagani, E., Rossi, G.: Ontology-addressable contents in p2p networks. In: Proceedings of the WWW'03 Workshop on Semantics in Peer-to-Peer and Grid Computing. (2003)
4. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: An architecture for storing and querying rdf data and schema information (2001)
5. Broekstra, J., Kampman, A.: Serql: An rdf query and transformation language (2004) Submitted to the International Semantic Web Conference, ISWC 2004.
6. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for Internet applications. In: Proceedings of the ACM SIGCOMM '01. (2001)
7. Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S.: A scalable content-addressable network. In: Proc. of ACM SIGCOMM '01. (2001)
8. Nejdl, W., et al.: Super-peer-based routing and clustering strategies for rdf-based peer-to-peer networks. In: Proceedings of the Twelfth International World Wide Web Conference (WWW 2003), Budapest, Hungary (2003)
9. Tang, C., Xu, Z., Mahalingam, M.: pSearch: Information retrieval in structured overlays. In: ACM HotNets-I. (2002)
10. Haase, P., Siebes, R., van Harmelen, F.: Peer selection in peer-to-peer networks with semantic topologies. In: International Conference on Semantics of a Networked World: Semantics for Grid Databases, June 2004, Paris. (2004)
11. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. Transactions on Knowledge and Data Engineering **15** (2003) 871–882
12. Maedche, A., Staab, S.: Comparing ontologies - similarity measures and a comparison study. In: Proc. of EKAW-2002. (2002)
13. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. In: IEEE Transactions on Systems, Man and Cybernetics. (1989) 17–30
14. Ahlborn, B., Nejdl, W., Siberski, W.: OAI-P2P: A peer-to-peer network for open archives. In: Workshop on Distributed Computing Architectures for Digital Libraries - ICPP2002. (2002)
15. Aberer, K., Mauroux, P.C., Datta, A., Despotovic, Z., Hauswirth, M., Punceva, M., Schmidt, R.: P-Grid: a self-organizing structured p2p system. ACM SIGMOD Record **32** (2003) 29–33
16. Wertlen, R.: Dfn science-to-science: Peer-to-peer scientific research. In: Proceedings of the Terena Networking Conference (TNC 2003), Zagreb, Croatia (2003)
17. Tatarinov, I., Ives, Z., Madhavan, J., Halevy, A., Suciu, D., Dalvi, N., Dong, X., Kadiyska, Y., Miklau, G., Mork, P.: The piazza peer data management project. SIGMOD Record **32** (2003)
18. Tempich, C., Staab, S., Wranik, A.: REMINDIN': Semantic query routing in peer-to-peer networks based on social metaphors. In: Proc. of the 13th Int. World Wide Web Conference, WWW 2004. (2004)