SemreX: a Semantic Peer-to-Peer Scientific References Sharing System*

Hai Jin, Yijiao Yu
Cluster and Grid Computing Lab
Huazhong University of Science and Technology, Wuhan, 430074, China
E-mail: hjin@mail.hust.edu.cn

Abstract

The design and implementation of SemreX is studied in this paper, which is a semantic Peer-to-Peer (P2P) scientific references sharing system. Compared with other systems, SemreX extracts the bibliographic information from the PDF file directly, categorizes the paper into a sub-field of ACM Topic based on the semantic matching, and provides PDF file exchanging and comments of papers sharing in the P2P manner. The routing algorithms in the prototype system are discussed, and the heterogeneous reference formats are investigated, which is an obstacle needs to be cleared to improve the accuracy of extraction.

1. Introduction

Information retrieval, paper sharing and exchanging among different researchers and groups are common behaviors in academic community. Some systems have been widely used, such as *Science Citation Index* (SCI). Usually, researchers are interested to know which papers are the most influential in a specific research field, and how many times a paper are cited. Moreover, how to get these papers conveniently is also a hot issue. Citeseer, a well-known system about the scientific papers sharing, provides not only the citation information of each paper, but also the electronic copy of papers freely [1]. In a word, it becomes a trend that the scientific papers are shared and exchanged via the Internet, and the convenient searching and downloading software is expected.

With the birth of Napster, the traditional Client/Server information service paradigm is challenged by the increasingly popular P2P paradigm, especially in the music and movies sharing scenes. Similarly, hundreds of scientific papers are resided in each computer, and most of them have been read and commented by researchers. Why not share these files and exchange the personal comments, which are helpful to the other researchers? Compared with the centralized systems, P2P-based systems are of less network delay and stronger robustness. We search all kinds of scientific papers, including journal papers, proceeding papers, technical reports, and dissertations. However, the existing systems, such as IEEE Xplore, ACM Digital Library, do not support the diversity of references. It is a fact that most of the traditional information retrieval systems only provide the syntax-based information processing and querying due to the huge amount of files.

SemreX is a semantic scientific references sharing system developed by us, which offers the references sharing services in computer science field based on the P2P technology, semantic classification, and semantic querying and matching. We attempt to present the implementation techniques and obstacles during developing the prototype system in this paper.

The rest of the paper is organized as follows. Section 2 overviews the related works about file sharing application systems both with the Client/Server model and the distributed P2P paradigm. The software architecture of SemreX is presented in Section 3, and the design of the P2P communication layer of SemreX is discussed in detail. Section 4 demonstrates the prototype system of SemreX briefly, and the routing algorithms in SemreX are described in Section 5. Section 6 discusses the heterogeneous information extraction. Finally, some conclusions are drawn.

2. Related Works

We briefly overview some scientific reference sharing systems related to our work. From the perspective of network architecture, these systems can be categorized into the *Client/Server* model and the *Peer-to-Peer* paradigm.

CiteSeer, scholar.google, and SCI are the most popular centralized scientific citation systems. A common feature of them is that there is a powerful server in the application system, which can get the

* This work is supported by National Basic 973 Research Program of China under grant No.2003CB317003.

Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT/ICIW 2006)
0-7695-2522-9/06 $20.00 © 2006 IEEE

complete documents from some digital libraries and compute the citations information accurately [1]-[3]. Most of the statistics works of these systems focus on the syntax matching, but the semantic matching is not provided. The single server leads to the single failure point and the bottleneck of services. Furthermore, many papers are not published in journals and proceedings, which are shared as technical reports, and dissertations. Unfortunately, these files are difficult to be collected by the centralized systems.

Distributed search over the Internet is popular since the emergence of P2P files sharing applications, such as Gnutella, Kazaa and eDonkey. The issues about the unstructured P2P file sharing system Gnutella, the design and implementation technology, the virtual topology, the users' actions and the traffic are addressed in [4]-[10]. Although these P2P applications are successful in the file sharing sense, most of the files shared are about music and videos, such as *mp3* and *mpg*. These systems cannot meet the requirements of scientific references citations, because they do not care the contents of the files.

Bibster is the most similar system to SemreX, which is a semantics-based bibliographic Peer-to-Peer system [11]. It extracts the bibliographic information from Bibtext files, and provides the sharing functions among different peers. Bibster first utilizes ontology technology in the P2P bibliographic sharing systems. However, Bibster neither extracts the references records from the original scientific papers, nor provides the file sharing function.

Our system SemreX gets the references records from the original PDF files, so it less depends on other systems. SemreX aims at supporting semantic queries and matching, and sharing the PDF files and citation information in the Peer-to-Peer manner.

3. Software Architecture of SemreX

SemreX is a complicated software system, which includes bibliographic information extraction, semantic classification of papers, semantic representation, syntax and semantic query, and P2P communications. In this paper, we only give an overview of the implementation and the detailed methods will be discussed in dedicated papers respectively. All the software components presented in this section focus on the P2P communication layer.

3.1. The software components of SemreX

The abstract software components of SemreX are illustrated in Fig.1. SemreX tries to extract references records from the original scientific papers. Because most of the papers are represented in PDF format, *Local Data Source* is the collection of PDF files. For security reasons, users have to give special directory, in which the PDF files will be shared. Other PDF files outside this directory will not be accessed. Since most researchers have more than one research interest, the papers are stored in different directories based on their topics. SemreX can access PDF files in several specified directories and all their sub-directories.

With the measurement of Gnutella, free riding brings negative effects to Peer-to-Peer application [7]. To reduce those free riders, SemreX requires each peer to specify at least a directory for shared PDF files. During the testing period, the number of sharing PDF files on each computer ranges from several PDF files to several hundreds. That indicates that the free riding phenomenon also exists in SemreX, which should be considered in the future development.

Controller coordinates all software components in SemreX working together.

Semantic Metadata Encapsulation accesses all shared PDF files, and extracts the title, author, keywords and references, and submits these heterogeneous information to the *Local Knowledge Repository Management*. Extracting the metadata from the original files is main difference between Bibster and SemreX. This component is difficult to be realized because of the heterogeneous references format. How to improve accuracy of extraction is an important issue in the design and implementation of this module.

The metadata are organized, queried and matched in the semantic method. *Local Knowledge Repository Management* performs these functions. The metadata is not managed by the relational database management system. However, they are represented in the RDF format and managed with a semantic database management system *SeSame*. Each paper about computer science is categorized into a special type in ACM Topic. With the ontology, we can get the similar papers of a specific topic. The semantic classification of a paper is only based on the references cited. In the future, it will use all information including the contents.

SemreX allows researchers to add their private comments about each paper and categorize papers into a specific sub-field of ACM Topic through the GUI.

Proceedings of the Advanced International Conference on Telecommunications and
International Conference on Internet and Web Applications and Services (AICT/ICIW 2006)
0-7695-2522-9/06 $20.00 © 2006 IEEE

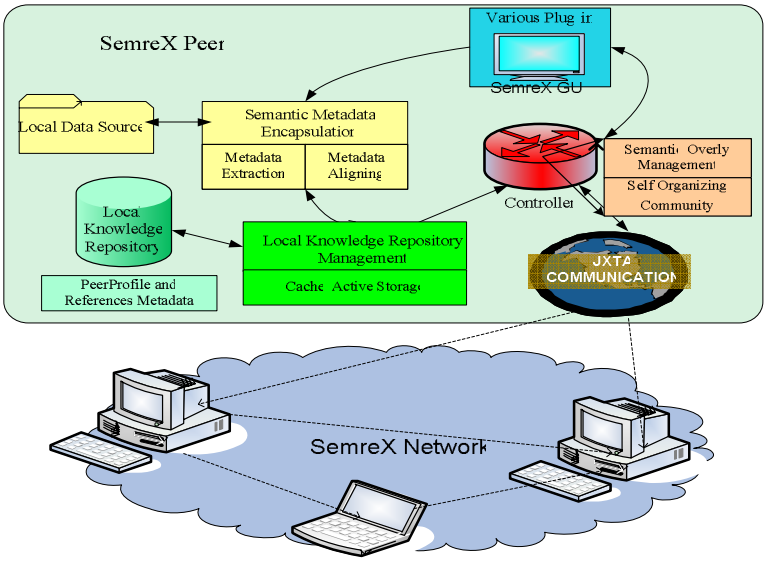

Figure 1 Software components in SemreX

These comments are stored in database, shared by all peers and regarded as the most important ranking metric of the scientific papers. The classification expertise from human is prior to the automatic classification results.

3.2. The P2P communication layer

In the prototype system, JXTA is utilized to build the P2P communication layer. JXTA is a popular open P2P protocol and software package [12], and some P2P application systems are built with JXTA [13]-[14]. The implementation of the P2P communication layer becomes easier by using JXTA. Part of classes and interfaces are listed.

LocalPeer listens to *InputPipe* and accepts query requests and query results. When receiving a message, it extracts the message type, fires events to notify other modules of SemreX. The communications among different modules are based on event-driven programming, and the operations related to the communication layer are asynchronous.

PeerDiscovery monitors the changes of the peer group. When a new peer joins, it notifies the human-machine interface of adding description information about the new peer in the peer list. If a peer leaves, the peer information will be removed. Each active peer in the peer group is recorded. All the peer entities in the SemreX group are encapsulated in *PeerEntry*.

Figure 2 shows the classes about data exchanging between each peer and the events between different software modules in a peer. All the exchanging data between peers are encapsulated into messages.

Due to the unreliable and unstable network environment, the delays of network operations are various, all operations related to the P2P communication layer are asynchronous. For example, when a peer wants to query references from other peers, it only sends a query request, and needs not wait for the response. Event-driven method is utilized in SemreX, and the human-machine interface is fired by *SemreXQueryResultEvent* to show the query results immediately. When *PeerDiscovery* discovers a new peer, it fires a *PeerEntryChangeEvent*. When a *SemreXQueryRequest* message is received, a *SemreXQueryEvent* is fired.

A rendezvous peer is specified, which runs in a computer with public IP address. Other peers run on computers with either public or private address behind of the NAT or firewall. The testing results show that JXTA is able to ensure the messages to pass firewalls.

4. The Prototype System

SemreX has been implemented successfully, and a screenshot is illustrated in Fig. 3. Although some processing parts of software components are not directly shown in GUI, we describe them from Fig. 3.

The down-left sub-window illustrates the ontology, ACM Topic, as a tree. Each paper is mapped to a specific node in the tree, which is the semantic classification result by the *Semantic Metadata Encapsulation* component.

The top-right component shows the query results, and the data are from the *SemreXQueryResult* message received by *LocalPeer*. SemreX users can define the searching scope with the top-left component in Fig.3.

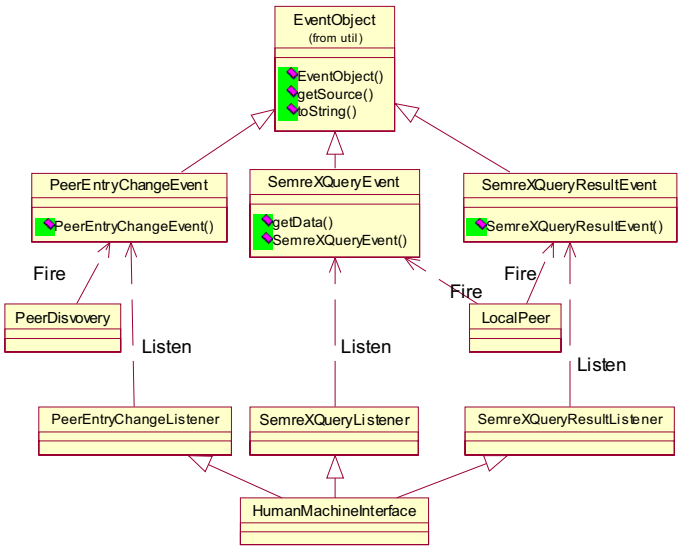

Figure 2 The events in the P2P communications layer in SemreX

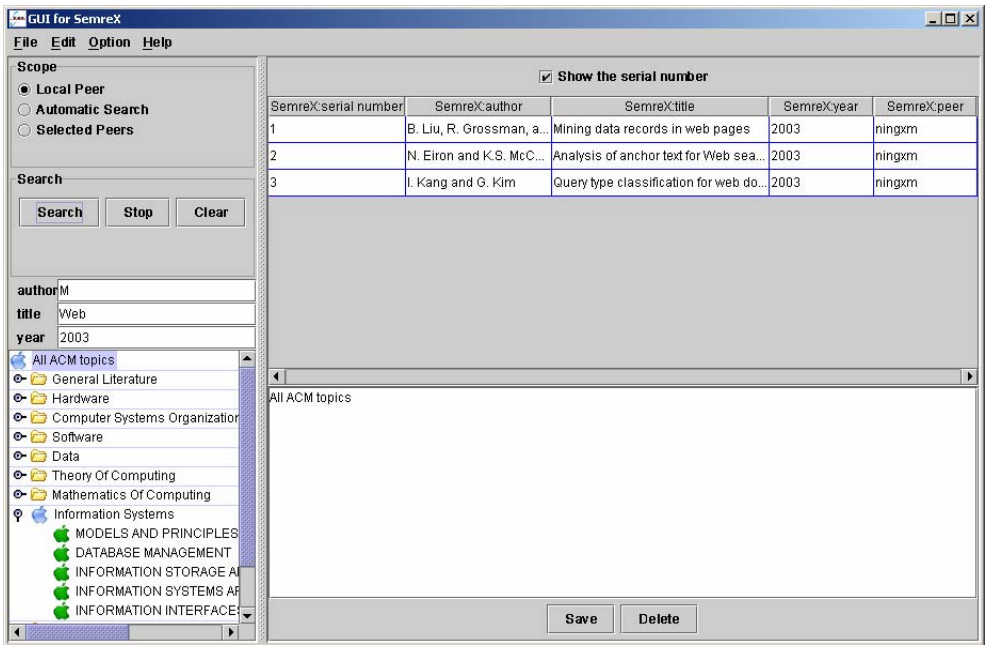

Figure 3 A screenshot of the SemreX system

If the *Selected Peers* is ticked, the information in *PeerEntry* will be shown, including *PeerName* and *PeerUUID*. Furthermore, when the ***Search*** button is clicked, *sendQuery* methods of the *LocalPeer* will be executed. The testing shows that the P2P communication layer with JXTA meets the system requirements of SemreX.

Communication performance is an important issue of the P2P application, and some related works have been done in [15]. At present, the communication layer of SemreX is powerful enough to support the references sharing. This may be contributed to the following reasons. First, the current P2P communication is implemented with the simple pipe services provided by JXTA, which cost less system resources. Second, the frequency of the exchanging operations among peers is not very high, because researchers are unlikely to search files many times in an hour. The performance of P2P communication layer in SermreX satisfies the proposed requirements.

5. Routing Algorithms in SemreX

Routing algorithms and routing protocols are key

issues of P2P application systems, which impact the efficiency directly. Some routing algorithms and protocols have been proposed in the file sharing systems, and the performances are reported in [15]-[16]. In the SemreX prototype design, JXTA is utilized as the P2P communication protocol, however, in the next version of SemreX the semantic-based routing algorithms will be employed.

5.1. The routing in the prototype system

As there is no global index of the PDF files and the references in SemreX, it can be regarded as an unstructured P2P network. Due to the medium number of users, unicast, multicast and broadcast models are provided. With JXTA, the peer discovery, join and leaving processes are transparent and users can see all the peers on line.

When a peer issues a query, it selects a peer, some peers, and all peers, and sends the query requests. From Fig. 3, query in the local peer is allowed by SemreX, which provides the information and documents management of the scientific papers in the desktop. Although the low-level routing processes are fulfilled by JXTA, the formats of the SemreX still need to be defined. For example, peers should give the type of the message, which tells how to do with the incoming message. The detailed tags of message defined in SemreX are defined.

SemreXMessageTypeTag: the type of the message. Two types are defined: *Query* and *QueryResult*.

SourcePeerNameTag: the name of the peer who sends the message. When the type of sending message is *QueryResult*, this data will be extracted and illustrated in the right column in Fig. 3.

DestinationPeeNameTag: the name of the destination peer.

DataTag: all the date information is encapsulated in this element in the self-defined XML formats. If the message is a *Query*, the data is the query field, such as the author name or the subjects of the papers. When it is a *QueryResult*, the data includes the list of references records, and all the information will be displayed in the top-right window in Fig. 3.

5.2. The semantic routing algorithm

The routing algorithms in semantic overlay are proposed in [17]. We are now investigating the routing algorithms based on the semantic similarity. Since all the PDF files are categorized into a sub-field of the ACM Topic, we categorize each peer into one or several sub-field of ACM Topic according to the cluster of the PDF files in each peer.

The basic semantic routing algorithm of SemreX is that the routing path only covers the similar peers in the overlay network. For example, there are five neighbor peers of peer *A*, marked as *B*, *C*, *D*, *E* and *F* respectively, and only *D* and *E* are very similar to *A*. With the semantic routing algorithm, just the most similar peers will be multicast. Then, only *D* and *E* will receive the query messages.

There are two key preconditions. First, the semantic description of each peer should be computed with the clustering algorithm. The second is how to compare semantic similarity between two semantic descriptions according to ACM Topic.

6. Heterogeneous Information Extraction

The various formats of references are an obstacle of information extraction and representation in SemreX. We find that the heterogeneity of bibliography originates from the different publishers and different types of papers. In this section, we only illustrate the heterogeneity of the formats.

First, the representations of sequential number of each record are different. Papers published by ACM and Elsevier Science include "[ ]", however, that of Springer Link does not.

Second, there are some differences about the representation of author's name. Usually, the first name is abbreviated in front of the surname. However, there are some exceptions. Furthermore, some papers give the full name of authors. SemreX supports query with the author name, and these differences bring troubles to realize this function. For example, it is not clear whether the bibliographic records with *H. Jin* should be returned when *Hai Jin* is queried. Moreover, most of papers are co-authored with several authors. The full name list of authors is given in some papers while some only partially list the authors by using *et al*. If an author's name is omitted, the results of query with name will not be correct.

Third, the representations of the paper title are different. In the IEEE case, the paper title is included by "", but others not.

Fourth, the page numbers are recorded in different manners, and even some papers do not have at all. If they are not given, the related items in the database will be assigned null, which leads to some inconsistency in database.

The sources of the papers include journals, proceedings, technical reports, standards or protocols, dissertations, and even the web pages on line. The description of journal name is similar and easy to be extracted, while the description of the volume and issue are different. In ACM and IEEE format, they are

depicted as *Vol.X, No.X*. However, *No.* is replaced by *Issue* in some papers. Elsevier adopts another style such as *20(8)*.

The diversity of proceedings is also an obstacle of information extraction. Some papers give the completed name of conference; however, some only use the abbreviations, such as the *INFOCOM*. The date and the location of the conferences are given in some papers, while others not.

7. Conclusions

The initial design and implementation of SemreX, a semantic scientific references sharing P2P system based on JXTA, is presented in this paper. The main software components, especially the P2P communication layer, are illustrated and the obstacles about the heterogeneous reference formats are discussed. The routing algorithms in the prototype system are described. Through the prototype system design, some experiences about the semantic P2P system are obtained, which is helpful to the future development of SemreX.

References

[1] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing", *IEEE Computer*, Vol.32, No.6, 1999, pp.67-71.

[2] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Networks*, Vol.30, No.1-7, 1998, pp.107-117.

[3] L. A. Barroso, J. Dean, and U. Hölzle, "Web Search for a Planet: The Google Cluster Architecture", *IEEE Micro*, Vol.23, No.2, 2003, pp.22-28.

[4] S. Waterhouse, D. M. Doolin, G. Kan, and Y. Faybishenko, "Distributed Search in P2P Networks", *IEEE Internet Computing,* Vol.6, No.31, 2002, pp.68-72.

[5] M. Castro, M. Costa and A. Rowstron, "Should we build Gnutella on a structured overlay?", *ACM SIGCOMM Computer Communications Review*, Vol.34, No.1, 2004, pp.131-136.

[6] E. Adar and B. A. Huberman, "Free Riding on Gnutella", *First Monday*, Vol.5, No.10, 2000.

[7] M. Ripeanu, A. Lamnitchi, and I. Foster, "Mapping the Gnutella Network", *IEEE Internet Computing*, Vol.6, No.1, 2002, pp.50-57.

[8] W. Wang, H. Chang, A. Zeitoun, and S. Jamin, "Characterizing Guarded Hosts in Peer-to-Peer File Sharing Systems", *Proceeding of Globecom'2004,* pp.1539-1543.

[9] D. Stutzbacn and R. Rejaie, "Characterizing the Two Tier Gnutella Topology", *Proceeding of SIGMETRICS'2005*, pp.402-403.

[10] D. Hughes, G. Coulson, and J. Walkerdine, "Free Riding on Gnutella Revisited; The Bell Tolls?", *IEEE Distributed Systems Online*, Vol.6, No.6, 2005.

[11] P. Haase, J. Broekstra, M. Ehrig, M. Menken, P. Mika, M. Plechawski, P. Pyszlak, B. Schnizler, R. Siebes, S. Staab, and C. Tempich, "Bibster – A Semantics-Based Bibliographic Peer-to-Peer System", *Proceedings of the Third International Semantic Web Conference (ISWC2004),* 2004, pp.122-136.

[12] L. Gong, "JXTA: A Network Programming Environment", *IEEE Internet Computing*, Vol.5, No.3, 2001, pp.88-95.

[13] A. Sanna, C. Zunino, and L. Ciminiera, "A distributed JXTA-based architecture for searching and retrieving solar data", *Future Generation Computer Systems*, Vol.21, No.3, 2005, pp.349-359.

[14] E. Halepovic and R. Deters, "Building a P2P Forum System with JXTA", *Proceedings of the Second International Conference on Peer-to-Peer Computing (P2P'02)*, 2002, pp.41-48.

[15] M. Portmann and A. Seneviratne, "Cost-Effective Broadcast for fully decentralized Peer-to-Peer Networks", *Computer Communications*, Vol.26, 2003, pp.1159-1167.

[16] E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim, "A Survey and Comparison of Peer-to-Peer Overlay Network Schemes", *IEEE Communications Survey and Tutorial*, March 2004, pp.1-22.

[17] C. Tang, Z. Xu, and S. Dwarkads, "Peer-to-Peer Information Retrieval Using Self-Organizing Semantic Overlay Networks", *Proceedings of SIGCOMM2003*, pp.175-186.