SemreX: Towards Large-Scale Literature Information Retrieval and Browsing with Semantic Association*

Xiaomin Ning, Hai Jin, Hao Wu
Cluster and Grid Computing Lab
Huazhong University of Science and Technology, Wuhan, 430074, China
Email: hjin@hust.edu.cn

Abstract

Access to scientific literature information is a very important, as well as time-consuming daily work for scientific researchers. Current methods of retrieval are usually limited to keyword-based searching using information retrieval techniques. In this paper, we present SemreX which implements efficient large-scale literature retrieval and browsing with a single access point based on semantic web technologies. The concept of Semantic Association is proposed to reveal explicit or implicit relationships between semantic entities, combining with the ontology-based information visualization technique so as to facilitate researchers retrieving semantically relevant information, as well as context relationships which can capture user's current search intentions while preserving an overall picture of scientific knowledge.

1. Introduction

Most users find information on the Web by browsing the information space or through the use of search engines, such as Google, Yahoo, Altavista. For browsing, information is usually hierarchically cataloged by topic and sorted by date, author, publication, etc. The user of browsing the information space obtains a navigable structure and makes own decision to the resource level intended. However, a hierarchical category system requires a high degree of pre-editing and also is too rigid for providing flexible views to users. Full-text search engines have gained the highest pervasiveness. Search engines deliver ranked results to users based on keywords or key phrases. However, none of these approaches can grasp users' intentions or information goals. Furthermore, these approaches can not reveal relevance or context relationships within documents.

Information retrieval methods based on semantic web [1] techniques promise advantages compared to the above conventional approaches. The semantic web is an extension of the current Web, based on the idea of exchanging information with explicit, formal and machine-accessible descriptions of meaning. Substantial efforts have been devoted to the development of the semantic web, including standards and recommendations [21, 22, 23] from the *World Wide Web Consortium* (W3C) and applications supported by the Semantic Web Community [20]. Documents or HTML contents on current web are rarely available in structured format and hold hardly any meta-information regarding the contents. The semantic web contains resources corresponding not just to objects (e.g., texts, images, audio, people, places, organizations) but also to relationships between objects [2]. So the semantic web is not a Web of documents, but a Web of relations between resources denoting real world objects, such as people, places and events.

For scientific researchers, access to scientific literatures is a very important daily work, including full-text searching and metadata (e.g., authors, title, abstract, publication, date) searching. Additionally, they are also interested in semantic relationships between metadata within documents, such as co-authors analysis, citation relationships, co-citation (two papers cited by the same paper) analysis, co-reference (two papers citing the same paper) analysis, relevant documents, similar documents. Furthermore, it is of significance for researchers to retrieve semantically relevant information and context relationships capturing user's current search intentions while preserving an overall picture of scientific knowledge, so as not to get "lost in hyperspace" [3] through a single access point other than merely browse a rigid hierarchical category or resort to combining numerous separate literature repositories, such as IEEE Electronic

* This paper is supported by the National 973 Key Basic Research Program under grant No.2003CB317003, and the Cultivation Fund of the Key Scientific and Technical Innovation Project, Ministry of Education of China under grant No.705034.

Library, ACM Digital Library, DBLP, CiteSeer, Google Scholar [24], to obtain the desired information.

We present SemreX, a system which addresses the above issues and implements semantic based large-scale literature retrieval and browsing. The next section gives motivation scenarios to show how the system facilitates access to scientific literatures. This paper also presents the concept of Semantic Association which describes explicit or implicit relationships between semantic entities (i.e., resources in semantic web). In SemreX, users can not only perform traditional full-text searching and metadata searching, but also can search or browse guided by following semantic associations to retrieve relevant information within large-scale literatures. In this system, the ontology-based information visualization technique is used to enable users to better express users' information goals and improve the interaction of users with entities. The visualization mainly focuses on entities and their relationships rather than merely on the ontological model, so it is very useful for literature retrieval.

The rest of the paper is organized as follows. Section 2 describes our motivation scenario. Section 3 reviews some related work. Section 4 presents the system architecture of SemreX. In Section 5, we propose the knowledge base, including the design of ontology schema and the definition of Semantic Association. Section 6 discusses the primary functionalities. Section 7 gives the preliminary implementation and results. Finally, conclusions and our future work are given in Section 8.

2. Scenario

We present the system for a computer science literature information retrieval environment which is expected to entail researches efficient literature information accessing with a single access point, including full-text searching, metadata searching and relevant information retrieval. Let us consider the following scenario supported by the system.

We assume that Alice, a Ph. D candidate in computer science program and major in distributed computing, now turns her research interest to information retrieval in distributed systems. However, she is almost completely unfamiliar with information retrieval except some experience on utilizing conventional search engines. She wants to quickly step into the research area of IR and acquire a sketch about the discipline, such as what IR means and what its goals are, what important articles and researchers are/were in IR, what primary conferences and journals are about IR, what major approaches or algorithms are involved, what other disciplines or research areas are closely related to IR (e.g., statistics, artificial intelligence, machine learning). Furthermore, she may prefer to know whether other researchers are taking part in the same research direction as her or how their research progress is or what publications they have published, so as to uncover what motivated her research work and prevent duplicating already known research results. At present, combining numerous separate literatures repositories and searching through mountains of literatures to tackle with the above tasks are not impossible but very time-consuming and intricate. Furthermore, when it comes to a multidisciplinary field of study (e.g., IR in distributed systems), it is rather difficult to maintain an overview of what is going on. However, with the help of our system, these tasks can be accomplished conveniently. In SemreX, Alice first performs a conventional full-text searching with keywords "*information retrieval*". Ranked results of documents containing the keywords are returned to her. Statistical analysis about all these records is also given, including publisher information (e.g., authority conferences, journals, technical reports, thesis), author information (e.g., productive authors, most cited authors, co-author relationship), content information (e.g., most frequent terms in these collection). In addition, she can select one of the records and follows the ontology property "similar" to get other similar literatures in content or the ontology property "related" to get relevant literatures (e.g., citing/cited relationship or presented at the same conference). Finally, she refines current results by the keywords "*distributed system*" to retrieve all relevant literature information about her current research direction. Thus Alice may get literatures which do not contain the query string "*information retrieval*" or "*distributed system*" at all since they may be authority publications and referenced by other publications about "*information retrieval*" or "*distributed system*".

3. Related Work

CiteSeer [4] is a digital library that aims to improve the dissemination, retrieval, and accessibility of scientific literature in computer and information science. CiteSeer locates scientific articles on the web, extracts metadata information such as the citations, article title, authors, and performs full-text indexing and autonomous citation indexing. Though CiteSeer provides some metadata relationships within literatures, e.g., related documents, citation context, similar documents based on sentences, it mainly emphasizes on citation analysis and can not reveal other complex relationships, for example, if two papers have no co-citation or co-reference relationship at all, CiteSeer will

consider no relationship between them though they could still have a strong relationship between them if their citations intersect substantially. Furthermore, CiteSeer can not retrieve context relationships capturing user's current search intentions while keeping an overall picture of scientific knowledge.

Flink [18] is a presentation of the scientific work and social connectivity of semantic web researchers, in particular the community of researchers who have contributed their work to the *International Semantic Web Conference* (ISWC) series. Flink employs semantic technology for reasoning with personal information extracted from a number of electronic information sources including web pages, emails, publication archives and FOAF [19] profiles. The acquired knowledge is used for the purposes of social network analysis and for generating a web-based presentation of the community. In Flink, the information visualization technique is used to display the ontology of research topics and analyze the social network of semantic web researchers.

Similar to ours, the notion of Semantic Associations is proposed [10] as complex relationships between resource entities. These relationships capture both the connectivity of entities as well as similarity of entities based on a specific notion of similarity called ρ-isomorphism. To discover semantic associations between two entities, the different path sequences interconnecting the two entities can be found. However, since each instance usually belongs to multiple classes in a large RDF graph, numerous possible path sequences may exist thus the computation of path finding is overwhelming. Furthermore, if the RDF schema graph is a connected graph, all semantic entities will be semantically associated. It is necessary to give a metric to measure the degree of the semantic association which is not given in [10].

As the semantic web is an extension of the current Web, it is supposed to grow with much more huge amount of data. Information visualization can provide functionalities to make the information more accessible and improve their usability. However, the use of information visualization techniques combining with semantic web standards is still mainly limited to research applications, with an exception of Aduna AutoFocus [5, 17]. AutoFocus is a commercial desktop search application that applies both information visualization (e.g., Cluster Map) and semantic web technologies (e.g., RDF, Sesame [16]). Though AutoFocus can combine conventional keyword search with various other means to browse and explore the information space, the relationship between documents is still mainly based on keywords and the metadata is very simple, just including documents location, date, file type, etc. So it can not uncover more complex relationships within documents.

4. System Overview and Architecture

The primary principle of the system is to extract the metadata (including article authors, title, abstract, year, publication name, references list) from large amount of computer science literatures and BibTeX bibliographies, aggregate (including duplication detection, citation context) them into the literature knowledge base. To provide full-text searching and other complex relationships within literature information, the system also supports full-text indexing and relationships between literatures (e.g., similar documents, related documents). The system architecture is depicted in Figure 1.

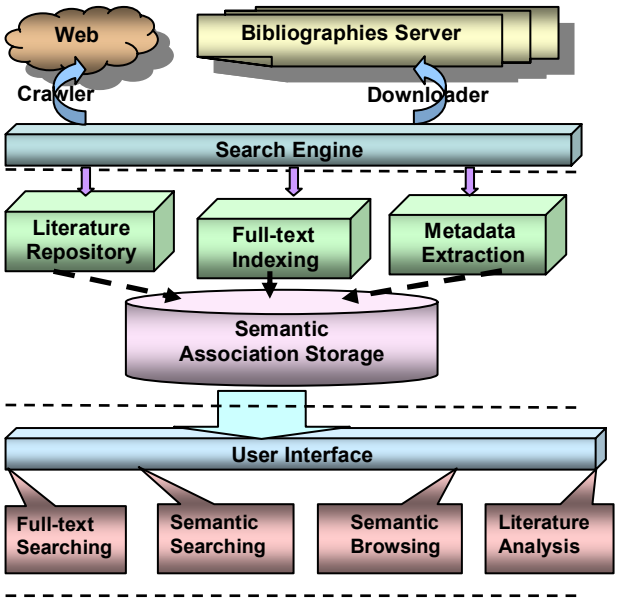


Figure 1. The system architecture

The system is separated into three layers: data sources layer, semantic association storage layer and user interface layer. In the data sources layer, we locate large amount of computer science literature sources freely available on the web, as well as BibTeX files from bibliography repositories (e.g., DBLP XML records [26], the Collection of Computer Science Bibliographies [27]) through search engines. After literature sources are gathered and stored in the literature repository, the system performs full-text indexing. Then metadata extraction module extracts the metadata and stores them according to semantic context within literatures. The system supports four

functionalities: full-text searching, semantic searching, semantic browsing, and literature analysis.

5. Knowledge Base

5.1 The Proposed Ontology

Ontologies are metadata that provide a controlled vocabulary of terms, where each term is defined explicitly so that it is machine-understandable. In general, an ontology is "an explicit specification of a conceptualization" [7]. Ontology schema can be comprehensive or lightweight. As [8, 9] indicate that lightweight ontologies are likely to become more important; not only are they cheaper to build and maintain, but they are more likely to be available off the shelf. So we tend to lightweight ontologies which describe the conceptualization of computer science literature.

Figure 2 demonstrates the major RDF schema applied in SemreX. For simplicity, some classes and properties (such as editor, volume/issue number of article) are omitted in the figure.

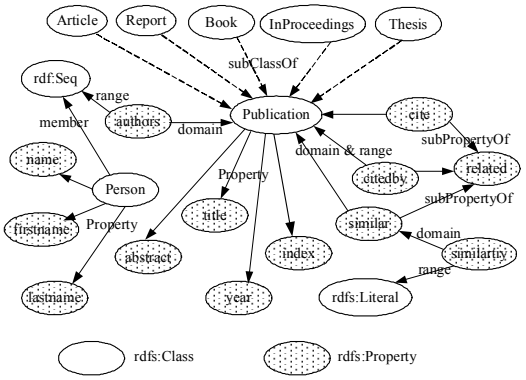


Figure 2. The ontological representation

Let us explain the above ontology schema. Each publication has not only metadata information, including title, authors, year, but also the relationships between them, including cite, citedby, similar, etc. The range of the property *authors* is a type of *rdf:Seq* which is the collection of all authors of the publication. The property *similar* represents the relationship of similar documents in content and *similarity* is the similarity degree which is very useful to retrieve ranked documents list in traditional IR. The properties *cite* and *citedby* describe the relationships that the publication cites others and is cited by others, respectively. As *cite*, *citedby* and *similar* all describe the relevance between publications, they are set as sub-properties of the property *related*, thus if the user wants to search publications related to the publication *P*, it is preferred to return the three sub-properties (i.e., *cite*, *citedby* and *similar*). Inference is also supported by the system so that other relationships can be inferred by or composed of the above direct binary relationships. For example, though the co-citation property of any two earlier documents in the bibliographic references by a later third document is not given, we can deduce it by querying whether two publications are cited by the same publication via *citedby*.

Here we briefly explain the property *index* in Figure 2. For the consideration of retrieval efficiency, we store the full-text literatures in the file system and ontologies in Sesame. The *index* property gives the location information in the file system. If a user is interested in the original literature, he can follow the *index* property to download it.

Figure 3 shows a small portion of the ontology model depicting our application.

```
<rdf:Description rdf:about="urn://grid.hust.edu.cn/Literature#392">
    <rdf:type rdf:resource=
          "http://grid.hust.edu.cn/Literature#InProceedings"/>
    <SN:title>Efficiently Ordering Query Plans for Data Integration
    </SN:title>
     <SN:authors rdf:resource=
          "urn://grid.hust.edu.cn/Literature#852657e85b38f72b776e"/>
     <SN:year>2002</SN:year>
     <SN:source>http://www.cs.washington.edu/homes/anhai/papers/
          ijcai99-workshop.ps</SN:source>
     <SN:presentedin rdf:resource=
      "urn://grid.hust.edu.cn/Literature#54c224871ea3466dcb0a221"/>
        <SN:abstract> We describe Streamer, the query-reformulation
             component of a data integration system. Given a utility
             measure and a user query, Streamer uses abstraction-based
             refinement planning and exploits information on plan
             independence to produce, in decreasing order of utility,
             a set of plans that access data sources to obtain answers
             to the query. We then focus on plan coverage as an
             important utility measure....</SN:abstract>
    <SN:cite rdf:resource="urn://grid.hust.edu.cn/Literature#106712"/>
    <SN:cite rdf:resource="urn://grid.hust.edu.cn/Literature#267886"/>
    <SN:cite rdf:resource="urn://grid.hust.edu.cn/Literature#121031"/>
    <SN:similar rdf:resource=
      "urn://grid.hust.edu.cn/Literature#57037c3bcc086b174d42a3e"/>
    <SN:similar rdf:resource=
       "urn://grid.hust.edu.cn/Literature#dee4f02635fde8be456e"/>
    </rdf:Description>
……
<rdf:Description rdf:about=
    "urn://grid.hust.edu.cn/Literature#54c224871ea3466dcb0a221">
    <rdf:type rdf:resource=
          "http://grid.hust.edu.cn/semantic/Literature#Conference"/>
    <SN:conferencename>IJCAI</SN:conferencename>
    </rdf:Description>
```

Figure 3. A portion of the ontology model

5.2 Semantic Information Extraction

The semantic web builds on contents to be described semantically via ontologies and metadata conforming to these ontologies. The quality of metadata extracted is of significant importance because the metadata will

be used to construct semantic context relationships. To tackle with large amount of literature information, automatic metadata extraction is indispensable. Automatic metadata extraction usually applies the information extraction [12] technique which is a low-cost approach to natural language processing using the finite-state automata technology to extract specific noun sets and information matching specific syntax and semantic templates. However, in our system, as the data sources to be processed are usually semi-structured scientific literatures (though with many different formats, e.g., IEEE, ACM, LNCS, Elsevier), the metadata extraction is more like a *parser*.

The extraction module performs the following extraction tasks:

- **Header of literature**. For each literature, the module extracts the header information, including the title, authors, abstract, publication date (if exists).
- **References/bibliographies list**. The reference list of each literature is extracted. For each reference entry in the list, the details (such as the title, authors, publication date, publisher information) are also parsed.

After the metadata is extracted, the semantic context relationships (i.e., *cite* and *citedby*) will be constructed. All the literature metadata pointing to the same entities (e.g., the same author) is integrated into the knowledge base. At present, we mainly use the rule-based parser and some heuristic techniques to retrieve the metadata of literatures. As machine learning methods can offer robust and adaptable automatic metadata extraction, we will combine *Support Vector Machine* [6, 15] method for metadata extraction to improve the accuracy.

5.3 Semantic Association Based Storage

The major goal of the system is to facilitate researchers retrieving semantically relevant information. The associations between semantic entities are usually very complex and implicit. We can simply divide associations into two types: direct associations and indirect associations. Figure 4 shows a small citation relationship sample with three *Publication* instances: *P1* (title: *The semantic web*), *P2* (title: *Three implementations of squishql, a simple rdf query language*) and *P3* (title: *TAP: A semantic web platform*). Direct associations are single binary relationships directly based on the ontology, for instance, *T. Berners-Lee, J. Hendler, O. Lassila* are *authors* of *P1*, *P3* cites *P1* and *P2*. Indirect associations are those connected by sequences of single binary relationships, for instance in Figure 4, *T. Berners-Lee* is not directly associated with *L. Miller*, however, their publications (i.e., *P1* and *P2*) are both cited by *P3* and there exists somewhat indirect association between *T. Berners-Lee* and *L. Miller*. If a user searches *L. Miller*, it may also be useful to provide him with the information about *T. Berners-Lee* (in fact, they are both interested in semantic web techniques.).

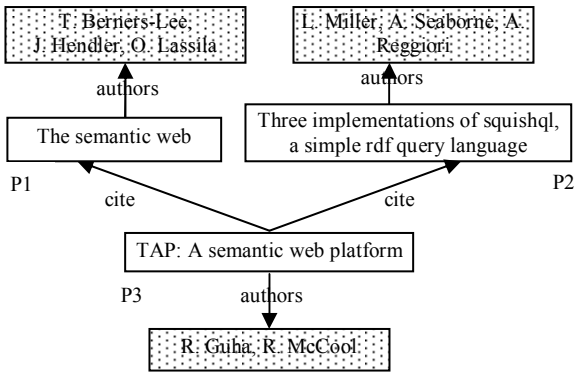


Figure 4. A citation relationship example

The *knowledge base* (KB) contains all semantic entities (instances of classes) and semantic relations (properties between semantic entities), besides ontology schemas. We generalize semantic relations and promote the concept of Semantic Association to denote the explicit or implicit relationships between semantic entities. Let the set of *semantic entities* be $E=\{e_1, e_2, ..., e_n\}$ and the set of *semantic relations* be $R=\{r_1, r_2, ..., r_m\}$. Given two entities e_i and e_j, only if they are connected through sequences of semantic relations when the RDF graph is taken as an undirected graph, there exists a semantic association between e_i and e_j. Connections can be found when they exist between any pair of concepts by applying a path finding algorithm. However, if the RDF schema graph is a connected graph, all semantic entities will be semantically associated. For example, as the literature ontology schema graph in Figure 2 is a connected graph, all authors and all publications will be semantically associated. So it is necessary to give a metric to measure the *degree* of the semantic association. The semantic association between any two entities can be ranked. At present, our work on this issue is in progress.

6. Semantic Retrieval and Browsing

6.1 Full-Text Literature Indexing

As our system is for literature information retrieval, supporting full-text search is still very important. In addition, full-text IR can be combined with the ontology-based retrieval method to deliver more

convenient functions. In SemreX, the full text is indexed using Lucene [25], an open source Java library for full text indexing and querying. Besides better performance it also gives us the ability to perform complex keyword queries, e.g. wildcard, phrase and proximity searches. We also calculate the similarity between literatures in content thus providing the retrieval of similar literatures denoted by the property *similar* in the ontology schema.

6.2 Semantic Retrieval

The most significant difference between our system and traditional retrieval systems is that semantic retrieval is supported. The system supports the following retrieval functionalities:

1. **Full-text searching**. Like traditional information retrieval systems, our system also supports complex keyword query, such as the combination of phrases, wildcard and proximity. The search result is the set of ranked documents satisfying the query.
2. **Metadata searching**. In this type of searching, the query is a boolean combination of entity names, e.g., finding the publication with authors="*T. Berners-Lee*" and title="*The Semantic Web*" and year="*2001*".
3. **Direct semantic relation based searching.** This type of queries involves using a specific relationship, for instance, finding all publications cited by the specified Article *A* via the property *citedby*.
4. **Semantic searching.** This type of searching is based on semantic association and may consist of sequences of semantic relations. In addition, semantic searching can be combined with the above three types of searching methods. For example, given a query string "*information retrieval*", the system will return the ranked documents according to the degree of semantic associations between the query and literatures, even if the publication does not contain the string since it may be referenced by other publications about "*information retrieval*" or may have been written by authors who have written other important "*information retrieval*" papers.

6.3 Semantic Browsing

Since the knowledge base contains large amount of instances and instances relationships, the system combines the benefits of the ontology-based browsing method and the visualization technique to facilitate researchers retrieving semantically relevant information, as well as context relationships which can capture user's current search intentions while preserving an overall picture of scientific knowledge. The ontology-based information visualization technique [11] is used to enable users to better express users' goals and improve the interaction of users with entities. In our system, the visualization mainly focuses on entities and their relationships rather than merely on the ontological model [13], this is very useful for literature retrieval. As the knowledge base to be visualized is very large (e.g., amounts of literatures, hundreds of authors analysis), we apply the Spring Embedder algorithm [14] to lay out so many elements according to the degree of semantic associations.

7. Preliminary Implementation

We have implemented simple semantic retrieval system, called SemreX, including full-text searching, metadata searching and direct semantic relation based searching. Semantic browsing is also supported. The data sources in our prototype are from CiteSeer Metadata [28], including the metadata archive *oai_citeseer.tar.gz* which is compliant with the Dublin Core standard with additional metadata fields (e.g., citation relationships, author affiliations, and author addresses) and the CiteSeer BibTex records. The CiteSeer Metadata covers literatures in the field of *Computer Science* and *Computer Technology* with about 800,000 publications and total approximately 400MB. To verify the feasibility and efficiency of our ideas, we use 50,000 publications in the prototype. The ontology storage is based on Sesame. The total number of RDF triples reaches 1,350,000. Lucene is used to index the full-text of the publications. Figure 5-7 give several interfaces to illustrate some functionalities of the prototype.

Although the prototype has implemented parts of our ideas, there is still much work to do. For example, the data set used in the prototype is very small and centralized compared with the information sources of the killer semantic web applications [8] to be distributed (both in terms of geography and ownership), to be heterogeneous and to contain real world data. With the size of the data set grows, more problems such as query performance, storage efficiency, will arise. Furthermore, more complex semantic searching and literature analysis functionalities should be supported.

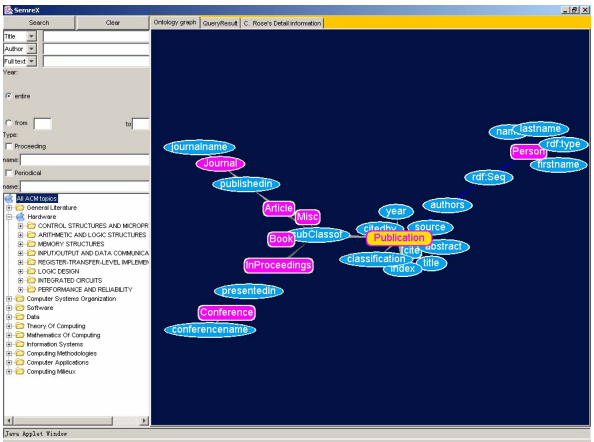


Figure 5. The ontological view and system interface

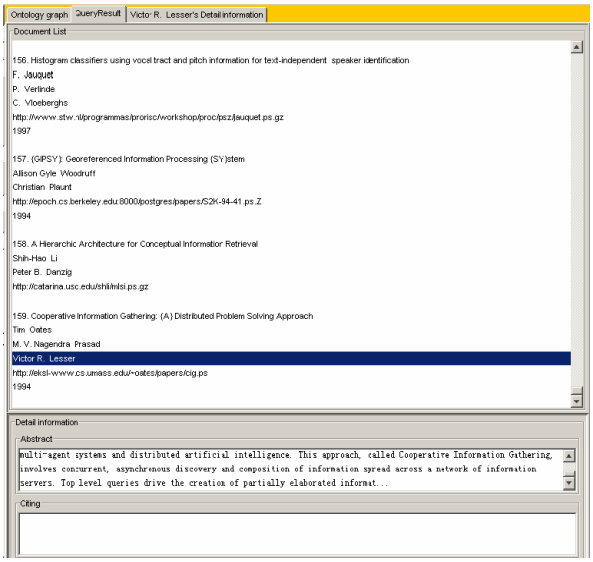
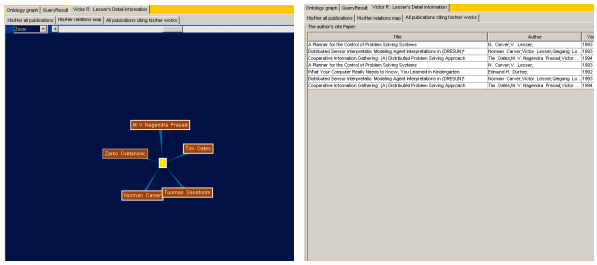


Figure 6. The query results interface



(a) Co-author relationship (b) Citation relationship

Figure 7. The relevant information on a specific author

8. Conclusion and Future Work

In this paper we present the SemreX system to implement efficient large-scale literature retrieval and browsing. We propose the concept of Semantic Association to describe explicit or implicit relationships between semantic entities. The ontology-based information visualization technique is used to facilitate researchers retrieving semantically relevant information. The system now supports three functionalities: full-text searching, simple semantic searching, and semantic browsing.

For our future work, we will give a metric to rank the degree of the semantic association, including assigning weights to the edges of semantic relationships or semantic entities. The performance of storage and retrieval is an important issue when the literatures grows to a very large scale (we expect approximately 1,000,000 literatures in the system). In addition, updating and keeping the knowledge base consistency is another significant concern to be tackled with.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web", *Scientific American*, May 2001.

[2] R. Guha and R. McCool, "TAP: A semantic web platform", *Computer Networks: The International Journal of Computer and Telecommunications Networking, Special Issue: The Semantic Web: An Evolution for a Revolution*, 42(5), 557-577, 2003.

[3] J. Nielsen, "The art of navigating through hypertext", *Communications of the ACM*, 33 (3), 297-310, 1990.

[4] S. Lawrence, C. Lee Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing", *IEEE Computer*, 32(6), 67-71, 1999.

[5] C. Fluit, "AutoFocus: semantic search for the desktop", In *Proceedings of the Ninth International Conference on Information Visualization*, 2005.

[6] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers", In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 1992.

[7] T. R. Gruber, "A translation approach to portable ontologies specifications", *Knowledge Acquisition*, 5(2), 199-220, 1993.

[8] H. Alani, Y. Kalfoglou, K. O'Hara, and N. Shadbolt, "Towards a killer app for the semantic web", In *Proceedings of the 4th International Semantic Web Conference*, 2005.

[9] M. Uschold and M. Gruninger, "Ontologies and semantics for seamless connectivity", *SIGMOD Record*, 33(4), 2004.

[10] K. Anyanwu and A. Sheth, "ρ-Queries: enabling querying for semantic associations on the semantic web", In *Proceedings of the 12th International World Wide Web Conference*, 2003.

[11] E. Hyvonen, S. Saarela, and K.Viljanen, "Application of ontology techniques to view-based semantic search and

browsing", In *Proceedings of the 1st European Semantic Web Symposium*, 2004.
[12] D. E. Appelt and D. Israel, „Introduction to information extraction technology", *IJCAI-99 Tutorial*, 1999.
[13] H. Alani, "TGVizTab: An ontology visualization extension for protégé", In *Proceedings of Knowledge Capture, Workshop on Visualizing Information in Knowledge Engineering*, 2003.
[14] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement", *Software – Practice and Experience*, 21(11), 1129-1164, 1991.
[15] H. Han, C. Giles, E. Manavoglu, H. Zha, and Z. Zhang, "Automatic document metadata extraction using support vector machines", In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 2003.
[16] J. Broekstra, A. Kampman, and F. van Harmelen, "Sesame: A generic architecture for storing and querying RDF and RDF schema", In *Proceedings of the 1st International Semantic Web Conference*, 2002.
[17] Aduna AutoFocus. http://aduna.biz/.
[18] Flink project. http://flink.semanticweb.org/.
[19] FOAF project. http://www.foaf-project.org/.
[20] The Semantic Web Community. http://www.semanticweb.org/.
[21] W3C Resource Description Framework (RDF). http://www.w3c.org/RDF/.
[22] RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/TR/rdf-schema/.
[23] W3C Web Ontology Language (OWL). http://www.w3.org/2004/OWL/.
[24] Google Scholar. http://scholar.google.com/.
[25] Apache Lucene. http://lucene.apache.org/.
[26] DBLP XML records. http://dblp.uni-trier.de/xml/.
[27] Collection of Computer Science Bibliographies. http://liinwww.ira.uka.de/bibliography/.
[28] CiteSeer Metadata. http://citeseer.ist.psu.edu/oai.html.