Collaborative, Distributed Information Management and Retrieval Architecture for the Enterprises

Tony Abou-Assaleh
Oliver Baltzer
Chris Jordan
Philip O'Brien
Hathai Tanta-ngai

Technical Report CS-2005-13

August 2005

Faculty of Computer Science
6050 University Ave., Halifax, Nova Scotia, B3H 1W5, Canada

Collaborative, Distributed Information Management and Retrieval Architecture for the Enterprises

Tony Abou-Assaleh, Oliver Baltzer, Chris Jordan, Philip O'Brien, and Hathai Tanta-ngai
Dalhousie University
6050 University Ave.
Halifax, Nova Scotia, Canada
{taa,obaltzer,cjordan,pobrien,hathai}@cs.dal.ca

ABSTRACT

The volume of published, digital content on the Internet grows rapidly each year. Locating and obtaining papers relevant to a particular topic becomes difficult as the sources are dispersed extensively making tracking and collection a lengthy and highly involved task. A solution to this fragmentation of publications is to define a common architecture whereby publishing firms provide local citation graphs to the remainder of the system enabling all firms to construct a global citation graph. The requirements of the participants of such a system are numerous but are lead by the need of each body to remain sole maintainer and provider of a published work. In this paper we present an architecture to address and accommodate the requirements of all parties involved in such a system. We show how our distributed citation system architecture is beneficial to the publishing firms involved in the distributed system as well as end users of the system.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*Retrieval models*

General Terms

Design, Management

Keywords

Enterprise information retrieval, distributed retrieval, collaborative information networks, citation graphs

1. INTRODUCTION

The number and variety of research-oriented publications added to electronic libraries each year is growing at a staggering rate. Coordination and partnership is needed to tackle the inevitable resurgence of information overload. Storage, management, navigation, analysis, and delivery of such vast repositories require equal amounts of partnership and delegation to those parties providing the published papers.

Researchers need tools that allow back-tracking and cross-referencing of documents and citations. Such tools must address the immediate needs and activities of researchers while maintaining sub-conscious requirement standards such as quality of service, efficiency, accuracy and intuitiveness. Such issues include easy access to relevant content, notification of publication meta data and their distribution. Publishing firms require complementary support including maintenance of meta data and assurance of its correctness, distribution of meta data and complete documents, and knowledge of publications from other firms for relevant and accurate cross-referencing with papers published externally. Consider the following example. Papers published through the ACM Press have internal cross-referencing allowing hyperlink creation on web pages to other documents within the organization's data store. Any references to published works outside this body are not linked.

The cornerstone of research navigation and the activities mentioned here is the citation. Citations are important when conducting creditable research. A citation exists when a given document is referred to by another. The higher the citation count for a document, the more other documents are building upon the work it is discussing. This is particularly important when such citations are non-reflexive as it can be inferred that members in the particular research community are endorsing it as a piece of important research. Papers that have the highest hit count can be viewed as pivotal pieces of work. A reference is when a given document refers to another; the converse of a citation. References are particularly important for recently published papers as it allows researchers to quickly find current work being done in their area. When discussing citation graphs, links of the graph represent a citation in one direction and a reference in exactly the opposite direction.

To construct an environment where citations and references are utilized to their fullest potential and in a way that provides robustness, accuracy, and dependability, contributors to the environment need to collaborate. In this context, collaboration requires the participation of all major publishing groups. These groups, however, will not easily share the contents of articles due to the desire to remain the only governing body over those articles. What is needed is a compensatory solution endorsed by all firms and is equally beneficial to users of the content. Firms require balances in:

- *Autonomy.* the ability to retract participation and ensure the subsequent retraction of the documents provided. In addition, collaboration does not assume complete replication of data and thus autonomic bodies require that while they share the published material, they are the *only body* providing that material upon request.
- *Return on Investment.* if collaboration is to occur, any enterprise investing resources will expect a return on that investment. These bodies require affirmation of their control and profit over documents shared and distributed.
- *Protection of Intellectual Property Rights.* There is a notion of access control when it comes to distributing publications. Publishing firms generate revenue by selling access to their repositories. However, being too restrictive limits the scope of these published works. Access control has to both restrict user access to these repositories while promoting their contents.
- *Reputation.* Each firm carries a reputation in the research community. It is important that any such collaborative system maintains these reputations. Likewise, confidence in the identity and dignity of other parties must be sustained.

We propose an architecture for collaborative, distributed construction of global citation and reference models targeted at research publications in Computer Science. This architecture is designed to benefit enterprises, such as publishing firms and search engines, while showing that amelioration of service to users can be achieved by the effective implementation of a distributed system. The system targets the issues described here while offering improved service and availability to consumers of research knowledge and information.

The remainder of the paper is organized as follows. Section 2 overviews related work in digital publication and citation management. We specify the problem in section 3. Section 4 presents the proposed solution. We we conclude with a summary in section 5.

2. RELATED WORK

To provide context to the problem, outlined in the following section, we take a look at related work in building citation indices and document repositories. Many tools have been produced for this reason and thus dominate much of this discussion. We divide these solutions into two categories: *centralized* and *distributed* systems. We further elaborate on each by providing an example.

2.1 Centralized Systems

Centralized information and knowledge repositories dominate this market. Used both as research assistants and organizational tools, centralized repositories have illustrated their importance. Scalability and efficiency are among the greatest concerns in many of the large-scale centralized information repositories. Many such repositories are limited by the number of documents that can be effectively stored and linked as well as by the number of remote connections they can serve. Both limitations clearly impact efficiency. Some of these tools include CiteSeer[1], The New Zealand Digital Library (NZ-DL)[2], MathSciNet[3], and Die Universität Trier's DataBase systems and Logic Programming (DBLP) Computer Science Bibliography[4].

The major drawback of centralized systems is the political red tape that must be cut in order to enable coordination between the publishing firms. Each firm wants to have its publications promoted in the best possible manner. Proving objective is critical and can be difficult to achieve especially when there are a few major stakeholders that dominate the field. Providing a level playing field where new stakeholders can enter may be politically difficult to do in this model. An unbalanced playing field can result in some work being obscured from view.

More dynamic and large-scale needs emerge from growing enterprises and proportionately growing article repositories; these enterprises offer compelling reasons to develop new solutions to address the many drawbacks of centralized systems. Distributed systems have the potential to overcome these drawbacks.

2.2 Distributed Systems

Bibster[5] — an open source peer-to-peer (P2P) system for managing, searching and sharing bibliographic data constructed from BibTeX files — routes queries to peers in a network, thus providing fast, dynamic search facilities. Bibster is distributed but does not promote collaboration or provider-side access control of the distribution. It operates only on meta data, not entire texts, even where complete texts are available. Consequently, users are left only with a pointer to a paper and not the paper itself, even if that paper is openly accessible.

Google Scholar[6] is distributed in the network sense in that it distributes services across multiple servers for increased parallelism but is not necessarily distributed in the societal sense. It does encourage collaboration by allowing institutions such as libraries and educational institutions to "post" their document repositories to the Google servers. To retract participation in the system, however, requires a manual and explicit request to Google to stop serving the documents previously shared.

OverCite [17] uses a distributed hash table (DHT) to store a fraction of the complete index table at each participating node. It also uses a DHT to store complete document files. Hashtables are not always scalable when used alone. Hash tables rely on unique addresses within the table for storage of values (documents). When the number of documents added exceeds the size of the table new entries begin mapping to locations of existing values. The modulus value must change

[1] `http://citeseer.ist.psu.edu`
[2] `http://www.nzdl.org`
[3] `http://www.ams.org/mathscinet`
[4] `http://dblp.uni-trier.de`
[5] `http://bibster.semanticweb.org`
[6] `http://scholar.google.com`

dynamically to accommodate this circumstance. Alternatively, chained hash tables may be used which bear their own detrimental features. The goal of OverCite is to reduce lookup time, load-balance across multiple servers, and reduce the amount of data maintained at each node on average (i.e., each node maintains nearly the same amount of information) by evenly distributing the global document set. OverCite suggests that documents be stored at some participating node in the system. This approach may not be desirable by publishing firms who are not willing to have their documents managed and stored at external sites.

Kešelj and Cercone [9] propose a framework for peer-to-peer (P2P) research collaboration based on semantic web and a push/pull architecture — PPDN. Variable and multiple agents [8] are distributed across a network to provide facilities for access control, delivery, and query execution. The domain addressed by this framework bears semblance to the one detailed in this paper, but is set in a P2P environment. It does not neatly fit into the scope of this paper for the following reasons: (1) publishing firms are large and static thus overhead associated with P2P hosts joining and leaving a network and the necessary update-flooding that ensues is trivialized, (2) identification and authentication protocols associated with P2P networks (e.g., [10, 16]) herald unnecessary complexity and performance compromises. While promising, PPDN addresses the P2P aspect of cooperation and not the integration of global citation graphs.

2.3 Centralized Example: CiteSeer

To better understand centralized systems we present an example of one: CiteSeer.

CiteSeer [4, 7] is one of the most well known and widely used non-publishing sources of research material within the traditional fields of Computer Science and closely related fields[7]. It indexes a large corpus of Computer Science research papers allowing users to browse and search in a vast inter-connection of cross-referenced publications. One of its most notable and advocated features is the citation analysis that it provides. In this type of analysis a directed graph is constructed with the nodes as papers and the edges represented by citations; this type of graph is referred to as a citation graph.

Using a citation graph, to derive contextual information regarding a paper, becomes tangible. Results come from analysis of how a paper is related to others. Papers citing papers on Information Retrieval exclusively are most likely about Information Retrieval. Frequently cited papers are said to have a greater impact than others and are thought to be more important. It can be inferred that a paper that has a large number of references from a variety of different authors is a survey of a particular subject. These are some of attributes that CiteSeer mines from its constructed citation graphs.

CiteSeer crawls the World Wide Web looking for Computer Science research papers that it subsequently downloads and stores in a central, monolithic database. Upon capturing these documents, citation analyses are performed. Following complete document processing, users of the system are allowed to query the database in addition to browsing and mining the citation graph. There are some shortcomings with CiteSeer's centralized design. In addition to the system being constantly overloaded with user's requests, CiteSeer does not index papers whose copyright is owned by commercial publishing firms (including publications from the ACM and the IEEE).

[7]A partial list might include: computational linguistics, neuro-computing, and e-commerce

2.4 Distributed Example: Google Scholar

Similar to above, to show a distributed citation system example we present Google Scholar.

Google Scholar is much the same as CiteSeer but capitalizes on Brin and Page's (now modified and unpublished) PageRank algorithm [15, 5] to achieve increased accuracy over CiteSeer. It does not, however, provide access (even limited) to documents that are not freely available on the Internet, but instead caches only links to the document. Access control takes the form of website maintainers actively denying access to content on websites by deterring robots and other crawlers. Proactive denial is counter-intuitive, especially for publishing firms considering the issues of copyright and intellectual property rights. Access should be granted only when desired by providers of academic papers [2]. Such access mechanisms leaves control of published content in the hands of the enterprise maintaining and managing it.

The citation graph obtained by Google Scholar is limited by the papers and references Google can ascertain from crawling the Web. Also, caching references is sub-standard to caching complete papers. If Google Scholar were to partner with major publishing firms to grant access to documents and thus deliver complete texts to users, the service provided to those users would be enhanced. As we further describe in section 4.2, if Google joined a collaborative system with, for example, the ACM and IEEE, requests for papers that appear in the Google Scholar results, but are actually located at the remote sites, could be delivered to the user via the Google's interface while abiding by the access control policies of the ACM and IEEE.

3. PROBLEM STATEMENT

The previous sections provided a motivation for our work. We are now in a position to define the problem addressed by this paper. In the research community there are three major identifiable enterprises: *publishing firms*, *research laboratories*, and *end users*.

In performing or using research, the requirements of each are as follows:

Publishing Firms

- Control of document repository and ability to withdraw at any time without repercussion
- Generation of profit or stability of resources (e.g., non-profit organizations)
- Retained rights of intellectual property

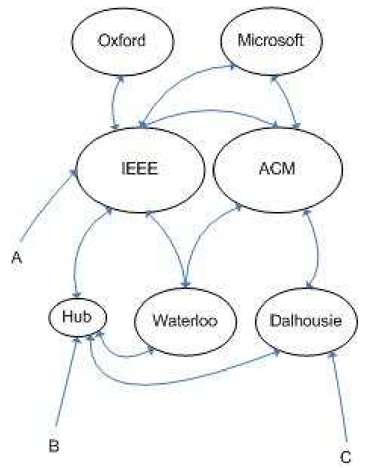

Figure 1: Node roles and redirection. Access is achieved through multiple entry points

- Strong and maintainable reputation in the research community
- Increased service to customers and members (increasing reputation)
- Reciprocity of resource investments (get more in return than was available before collaboration)

Laboratories

- Access to resources of importance (e.g., publications)
- Integrated and complete view of knowledge repository (e.g., no broken links in the citation graph)

End Users

- Access to research resources (e.g., publications, citations, and meta data)
- Single and multiple entry points (e.g., global and local search)
- Value-added service (e.g., analysis of the citation graph)

A system architecture is needed to accommodate all these requirements. What we propose next attempts to address each requirement in turn and provide a comprehensive solution that encourages the enterprises involved to participate in the collaborative, distributed system and realize the mutual benefit of doing so. By cooperating to provide a more complete citation graph index, each participating entity receives a return in the form of satisfaction from its members, interest from its customers and reciprocity from the other participants within the system.

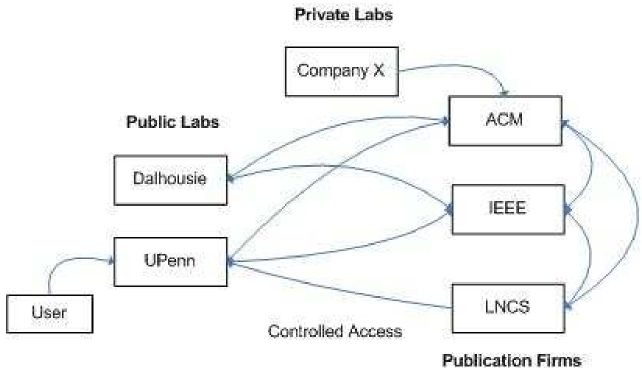

Figure 2: Distributed citation graph showing publishing firms, laboratories, and end users

4. DISTRIBUTED CITATION INDEX

The global citation graph is based on documents published by various institutions such as publishing firms, universities, corporations and individuals. In the proposed model each of the institutions provides a subgraph, or graph fragment, that contributes to the global citation graph. A graph fragment produced by an institution will have internal edges between documents that it publishes. In addition, this fragment will have external edges to documents in graph fragments belonging to other institutions. Thus the global graph can be viewed as an arrangement of all graph fragments connected by external edges.

The notion of primary and secondary authorities is modelled after the Domain Name Server standard [11, 12]. Primary and secondary domain name servers relay requests for host name address lookup to servers that are better able to service the request based on locale and resources. Load balance and timely response find synonymous meaning for these authorities.

A primary authority is defined as an institution providing a graph fragment based on its own document corpus. The authority will be responsible for that graph fragment and the fragment is given a unique identifier. A node in the global graph is a document and is uniquely associated with the graph fragment to which it belongs. This gives primary authorities full access control over distributed corpora.

Secondary authorities provide relief for primary authorities by sharing the load. Queries or other requests made to primary authorities or *hubs*, described below. Secondary authorities maintain a complete citation stack in the same manner as primary authorities. Alternatively, secondary authorities, like secondary DNSs, may maintain only portions of GCGs based on temporal clustering, access frequency, and so forth, delegated to them by their associated primary authority.

The end-user of a distributed citation index system is connected to the network through a mediating hub. Hubs are specialized nodes of a distributed system [13] that act as proxies for interpretation and redirection of queries to the global graph on behalf of the end-user and forward the queries to the appropriate authorities. Hubs are query dependant and will forward a query to a set of authorities, either primary or secondary, to ensure maximum coverage of the desired graph. In the case of a query over the complete CiteSeer graph (e.g., to determine the citation count of an article), the hub must query every graph fragment represented by either a primary or secondary authority.

To better understand this system configuration, consider Figure 1. There are multiple access-points to the distributed citation index system: directly through primary authorities (e.g., A goes to the IEEE Xplore website), through secondary authorities (e.g., user C using Dalhousie's proxy service to the ACM digital library) or through one of various hubs in the network (e.g., perhaps B connecting through a local ISP). If access is requested through A in Figure 1 for documents that may exist in both ACM and IEEE, the request may be redirected through the Hub. The hub provides relief for the primary authorities and requires only knowledge of the global citation graph in order to send portions of the requests to the primary authorities as required. Results are returned to A via the hub.

Every primary authority, secondary authority, and hub in the network maintains a complete list of all primary authority systems. If a new system joins the network it can obtain this list from any other system currently in the network. They can then query the primary authorities to provide information about their optional secondary authorities. As a result, the system scales simply as few nodes need to be notified when new participants join.

To reduce the network load, the hub can attempt to forward or route queries such that a minimum number of systems need be to involved in a search. This procedure is useful when a primary or secondary authority also serves as a secondary authority for some other graph fragment that needs to be queried. In this case the hub can pack queries over different fragments into one request negating the need to contact an additional authority. A detailed design and analysis of the query processing in the system will be addressed in an analysis of the proposed architecture. In addition to its query forwarding responsibilities a hub can also cache query responses and construct a cached fragment of the global graph for faster response to recurring queries. Mechanisms used to notify caches of changes to the graph structure will also be addressed during an analysis of the proposed architecture.

A response to a query commonly contains a set of documents linked through citations. Optionally, the primary or secondary authorities that have been queried can also recover Universal Resource Locators (URLs) for the full text of a document. Publishers of documents in a graph fragment are also the primary authority for that fragment; the publisher has absolute control over the URLs providing access to the full text of the document. Giving publishers this responsibility allows them to maintain their current licensing model for documents; they often grant access to documents based on IP-addresses for educational institutions and username/password for other subscribers. This is a course-grained approach to access control and inhibits effective exchange of documents and resources. Alternative access methods exist which provide more robust and dynamic access such as role based access control (RBAC) [18].

The external edges between graph fragments are exchanged among their corresponding primary authorities to validate edge endpoints. Primary authorities must register all incoming and outgoing external edges. To provide additional distribution, resource utilization and to support primary authorities with large document corpora, institutions can mirror the graph fragments provided by primary authorities; these institutions will be referred to as secondary authorities. The primary authority owning the fragment must authorize the mirroring and is responsible for disseminating changes to the graph fragment to its secondary authorities. Update notifications are also sent to other primary authorities whenever a graph fragment is altered thus allowing associated fragments to be updated.

We submit not to providing all documents to all persons at all times, but instead provide a common infrastructure for users to access published articles they would normally have access to without navigating to numerous sites. By making a request to LNCS, documents that are accessible by the requesting party at the ACM are also returned.

4.1 Case Scenario

Consider the diagram in Figure 2 of various research laboratories and publishing firms. The laboratories consist of two Universities and a private company. The publishing firms consist of the ACM, IEEE, and Springer's Lecture Notes in Computer Science (LNCS)[8]. All laboratories and firms run and maintain their own local citation graph system for their own purposes.

Suppose universities wish to share and exchange research openly. Company X has only access to a single database, the ACM library in this example. Company X does not wish to share its information as it is proprietary and sensitive in nature. laboratories have purchased access rights from publishing firms to access that data. They also have the option to share the citation analyses on their local digital repositories with each other. This sharing is beneficial to universities from the research perspective but not for Company X, from the business perspective. It is also in the best interest of the publishing firms to share citation analyses freely to construct the global citation graph ultimately enhancing the service provided and the search capabilities available to users.

Another benefit awarded to this approach is the integrity of document meta data. Reputation and social capital incentives ensure owners and providers of information maintain the data they provide thus sustaining a higher quality of data and information access, resulting in sustained access to the data by users – in other words, increased stickness [6] of the information repository.

4.2 Distributed Architecture Stack

Building from the scenario of Figure 2 we propose the architecture of Figure 3 to address the issues outline above.

Each participant of the distributed citation graph system

[8] `http://www.springer.de/comp/lncs/`

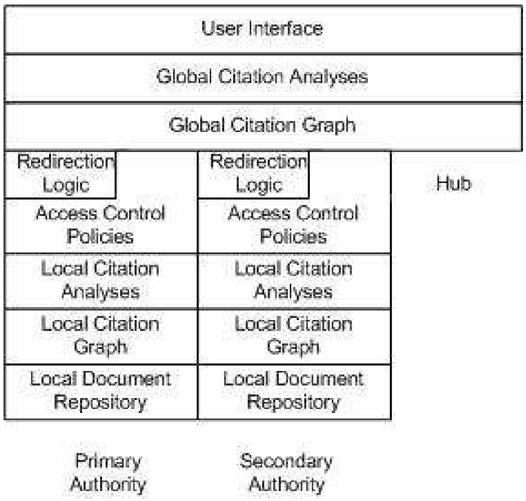

Figure 3: Distributed Citation System Architecture

maintains an independent, but logically overlapping, stack. At the foundation of this architecture lies a wide-spread and *virtually* shared global citation graph (GCG). Authorities participating in the system and contributing to the graph maintain their local citation graphs, and may optionally maintain a copy of the global graph. Queries on the system are first inspected at this layer but are executed and resolve matching sets later at the local citation graph layer.

From inspection of the GCGs out-going links for a given document, each publishing firm is able to redirect submitted requests for references, meta data, and documents to the appropriate maintainer of the local graph fragment within which the requested entity is located. Redirection employs a mapping between requested documents, local graphs and the network node holding that document. In this way, the GCG acts as an overlay to the physical network underlying the citation graph, similar to the technique used in multi-database management systems [14]. Remote users (e.g., corporate or academic laboratories, individuals) interface with the grid of participating nodes via a predetermined set of access points and remain oblivious to the redirection and retrieval methods employed. This multi-point access framework alleviates unnecessary searching or consideration from the user. For queries on local graphs, redirection is not required.

When the sites retaining the requested documents are found, queries and connections are relegated to those points and access privileges are resolved.

Both methods employ variable granular access to content. Entire repositories, groups of publications, areas of interest, temporal clusters (e.g., most recently viewed papers), individual papers, sections of papers, and meta data are examples of granular levels of content. Semantic Web [3] purports to address this issue by allowing access to content at arbitrarily refined levels. AC techniques are not described further. The Symposium on Access Control Models and Technologies [1] maintains an extensive library of proceedings on access control issues.

If access is granted, communication moves down the stack to the analytic layer of citation graph analyses. At this stage contextual information is gathered on the nature of publications: articles with a relatively large bibliography (high count on outgoing links in the GCG) may be surveys, those with a high count on incoming references are likely to be regarded as seminal publications in a field. This layer of the stack exists for contextual retrieval purposes. It is situated at this level as it relies on the LCG for its analysis but can be accessed only by authorized users (e.g., laboratories, individuals, companies).

The analytic layer is not, however, required to browse and navigate the local citation graph. This layer of the stack maintains the connected graph of those documents located at the site. The LCG contains one-ended links from documents in the LCG to known but unreachable documents outside the LCG. This state of the graph provides the catalyst when the site LCG is spliced with external LCGs to form or update the GCG.

Finally, below the LCG rests the local document repository. Processes operating on this layer update the adjacent two layers. Newly introduced documents are first processed and cross referenced with documents in the LCG, then introduced as new nodes appropriately. References to documents outside the LCG are left single-ended. The GCG is then updated and single-ended links are fastened to exterior nodes (i.e., documents located at other hosts). Deleted or updated documents follow similar processes.

The key to this architecture is the exchange of local citation indices with other participants of the distributed citation system. Communication between the publishing firms occurs below the system stack. Remote firms are treated as specialized users, as in the architecture diagram of Figure 3 and must undergo the same types of access control authentication. Exchange of LCGs and parallel updating of firms' respective GCGs provides for real-time accuracy and integration of the enterprises.

5. CONCLUSIONS

The proposed architecture delegates maintenance and access control responsibilities to the parties providing research documents to the system. This allows publishers to maintain the citation graph of their own documents and provide external references to other publishers. Without restricting the quality of service provided to the research community, this approach has several advantages for individual publishers:

- The publisher can maintain its licensing model for their documents
- No third party requires access to the document corpus to construct the citation graph
- No third party requires access to the complete document index or citation graph
- While catering towards these individual interests, the system is expected to receive a high acceptance and significant support from publishers

Distributing a system such as CiteSeer in this manner will address the shortcomings of its current centralized design. By delegating access control to publishers and sharing graph fragments, this approach will promote construction of a complete global citation graph. Publishers will be more inclined to provide high quality meta data for their documents since it will increase their visibility thus attracting more readers. The distributed nature of the architecture allows for greater availability and increased robustness for users. We believe these reasons provide justification and motivation for pursuing this project further.

This paper has proposed an architecture to effectively distribute the global citation graph and document store observed in CiteSeer. The system is intended to catalyze the partnership between the major players of the computer science research community. Enterprises interested in contributing to a larger, more comprehensive vision of shared citation analysis and in improving the experience of its members would see great promise in this design.

Acknowledgements

We would like to thank Dr. Vlado Kešelj, Dr. Evangelos Milios, and Peter Spierenburg for their suggestions and remarks.

Tony Abou-Assaleh and Chris Jordan are supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) postgraduate scholarships.

6. REFERENCES

[1] Symposium on access control models and technologies. http://www.sacmat.org.

[2] H. Berghel. Responsible web caching. *Communications of the ACM*, 45(9):15–20, 2002.

[3] T. Berners-Lee. Semantic web. In *XML2000*. World Wide Web Consortium, December 2000.

[4] K. Bollacker, S.Lawrence, and C. L. Giles. Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Agents '98: 2nd International ACM Conference on Autonomous Agents*, page 116, 1998.

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

[6] A. A. Bush and A. Tiwana. Designing sticky knowledge networks. *Communications of the ACM*, 48(5):67–71, 2005.

[7] C. L. Giles, K. Bollacker, and S. Lawrence. Citeseer: An automatic citation indexing system. In *DL '98: 3rd ACM Conference on Digital Libraries*, pages 89–98, 1998.

[8] V. Kešelj. Multi-agent systems for internet information retrieval using natural language processing. Technical Report CS-98-24, University of Waterloo, 1998.

[9] V. Kešelj and N. Cercone. PPDN – a framework for peer-to-peer collaborative research network. In *WSS '04: Proceedings of The Second International Workshop on Web-based Support Systems*, September 2004. In conjunction with 2004 IEEE/WIC/ACM International Conference on Web Intelligence.

[10] S. Marti and H. Garcia-Molina. Limited reputation sharing in P2P systems. In *EC '04: Proceedings of the 5th ACM Conference on Electronic Commerce*, pages 91–101, New York, NY, USA, 2004. ACM Press.

[11] P. Mockapetris. Domain names - concepts and facilities. RFC 1034 (Standard), November 1987.

[12] P. Mockapetris. Domain names - implementation and specification. RFC 1035 (Standard), November 1987.

[13] J. C. Navas and M. Wynblatt. The network is the database: Data management for highly distributed systems. In *SIGMOD '01: Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pages 544–551, New York, NY, USA, 2001. ACM Press.

[14] M. T. Özsu and P. Valduriez. *Principles of Distributed Database Systems, 2nd Ed.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1999.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[16] N. Saxena, G. Tsudik, and J. H. Yi. Admission control in peer-to-peer: design and performance evaluation. In *SASN '03: Proceedings of the 1st ACM Workshop on Security of Ad Hoc and Sensor Networks*, pages 104–113, New York, NY, USA, 2003. ACM Press.

[17] J. Stribling, I. G. Councill, J. Li, M. F. Kaashoek, D. R. Karger, R. Morris, and S. Shenker. Overcite: A cooperative digital research library. In *IPTPS '05: Proceedings of the 4th International Workshop on Peer-to-Peer Systems*, Ithaca, NY, February 2005.

[18] M. Wilikens, S. Feriti, A. Sanna, and M. Masera. A context-related authorization and access control method based on RBAC. In *SACMAT '02: Proceedings of the Seventh ACM Symposium on Access Control Models and Technologies*, pages 117–124, New York, NY, USA, 2002. ACM Press.