STUDY OF USER QUALITY METRICS FOR METASEARCH RETRIEVAL RANKING

Summary: Emory University and Virginia Tech seek funding to conduct a research project to gather foundational data concerning user quality metrics for metasearch retrieval.

Introduction: New metasearch systems that support simultaneous search and retrieval of heterogeneous sources of information suffer from a lack of empirical data and models concerning what users want in combined retrieval systems. Metasearching systems, which may incorporate technologies such as Open Archives Initiative metadata harvesters, portals, catalogs, and web search engines, now have the capability to automatically aggregate and make searchable many fundamentally different types of information (e.g. library catalog records, digital library item-level metadata records, web pages, collection level digital descriptions, and archival finding aids). But while there is a significant body of research concerning user preferences and searching behaviors in segregated systems such as library catalogs and web search engines, virtually no real-world user studies exist for metasearch systems that bring jarringly different sorts of information together. Little is known about optimal interfaces or visualization methods. Moreover, virtually no digital library search engines provide the capability for explicitly tailoring retrieval rankings that reflect the desires of specific user communities relative to the varying quality and attributes of the underlying resources.

Research Activities: This research project will undertake a series of studies using production digital library services to determine what criteria underlie the preferences and assumptions of different groups of users regarding metasearch systems. We will augment existing open source search engines, adding algorithms and interfaces to handle custom search ranking metrics based on the attributes of resources and collections. Building upon this testbed, we will experimentally assess the reactions of users to different retrieval algorithms and different quality metric weightings. The statistical responses of users will be analyzed to theoretically model user quality metrics for metasearching systems of various types. Finally, a dataset of results will be made available publicly and results will be reported in the literature and through presentations.

Research Outcomes: This work will have three main outcomes for both the research community and practitioners:

1. Empirical data concerning user preferences, expectations, and other quality metrics for metasearching systems will be made available.
2. A theoretical model of user quality metrics for metasearch systems will be produced, which will fit the large amount of empirical data collected.
3. A digital library search engine capable of accepting explicit ranking algorithms for metasearching will be produced and made available as open source software for other research and practical use, which will be based on the theoretical model.

NARRATIVE

SECTION 1: NATIONAL IMPACT

The Advent of Metasearching

The landscape of information retrieval possibilities has dramatically changed in a short period of time in recent years. Many information types that were previously held in data silos or secured in intranets can now be pooled in discovery services based on techniques such as metadata harvesting and federated searching. The concept of *metasearching* has recently gained prominence. In this proposal "metasearching" is used in its broadest sense to refer to the variety of techniques that enable users to simultaneously search multiple bodies of information that have previously been searchable only through separate interfaces. Many commercial vendors and non-profit services have developed metasearch systems in an effort to provide improved discovery mechanisms to end-users. To date, such endeavors have frequently overlooked the many problems that end-users face. For example, users may be confused by a disorganized and overly large set of results, or have works of lower quality more highly ranked than those of higher quality. There are many gaps in knowledge about user's implicit expectations and perceptions of "quality" in retrieval rankings indexed by metasearch systems from heterogeneous bodies of information. These expectations and perceptions have yet to be addressed by the library and information science research communities.

Challenges of Metasearching Heterogeneous Information

With the advent of new protocols such as the Open Archives Initiative Protocol for Metadata Harvesting (hereafter abbreviated OAI-PMH) digital library services suddenly gained the ability to rapidly and automatically harvest enormous quantities of newly exposed metadata from information repositories around the world. However, due to the flexibility of OAI-PMH, and the large number of varying assumptions surrounding the creation of metadata records, the heterogeneous metadata thus harvested often has incongruous and incommensurable characteristics. These characteristics include size, granularity, extent of vetting, ratings, review, impact, bibliographic coupling, and many more. These complexities must be faced, since researchers engaged in the study of focused subject domains desire all relevant information indexed in search and discovery services, encouraging discovery services to harvest and unify as many relevant sources of information as possible.

Such services should help users face the serious problem of dealing with a large, unorganized, heterogenous mixture, with widely varying levels of quality (along a number of dimensions). In cases of homogenous metadata such as uniform bibliographic citations to items that are relatively similar (articles in an Abstracting and Indexing database, for example), both users and system designers have a reasonably clear understanding of the nature of the records that will be retrieved by the system, as well as a general sense that items of greater relevance (gauged by various quality metrics, such as word frequency or citation frequency) should be ranked higher in the retrieval set. But the information indexed in metasearching systems often violates such assumptions, being heterogeneous in terms of granularity and fundamental level of abstraction. In the most extreme cases of metasearching, both digitized content (web pages crawled from the Internet, for example), are brought together with metadata exposed via OAI data providers. Under such circumstances, there is little formalized understanding of what constitutes quality metrics for ranking retrieved records in result sets, or how to appropriately contextualize the results in terms of their nature and level of abstraction. A review of the literature reveals the dearth of empirical studies of such user quality metrics in metasearching.

Summary of Literature Review

An extensive literature review seeking empirical studies of user expectations concerning quality in metasearching of heterogeneous information was conducted as part of the MetaScholar Initiative. This search was conducted because several MetaScholar projects seek to develop focused scholarly portals, each indexing a variety of harvested metadata and content. [Halbert, 2003a] The surprising lack of such empirical studies was striking, and ultimately led to this proposal. The literature search is only summarized here, in the interests of space, but more details are provided in Appendix B of this proposal.

Until recently information science literature has been overwhelmingly based on the assumption of record homogeneity and has focused on text-based measures of relevance. Representative review works [Baeza-Yates, 1999; Witten, 2002] demonstrate the gap in basic empirical studies of user quality metrics for metasearch in the case of predominantly heterogeneous sources.

This theoretical gap has recently started to be addressed. Awareness of the general complexity and ambiguity surrounding the quality-of-data problem has been articulated in recent years in the proceedings of the International Conference on Information Quality. However, much of this research has focused on the needs of sophisticated researchers for high quality data, rather than the needs of the average individual searcher. Attempts to theoretically model data quality have been produced [Wand, 1996], but these have lacked empirical studies of actual users.

The Need for Better Understanding of User Quality Metrics

This project attempts to address the dearth of empirical studies concerning the problem described above. User expectations and implicit assumptions about the quality and desired ranking of retrieved records in metasearch result sets is fundamental to the task of designing and configuring metasearch systems in coming years. Without such empirical studies, there will continue to be a gap in our understanding of how such systems should function in order to best meet user needs. Without systematic study, this gap will gradually be filled through anecdotal complaints by users, if at all.

Research Outcomes

This work will have three main outcomes for both the research community and practitioners:

1. Empirical data concerning user preferences, expectations, and other quality metrics for metasearching systems will be made available.
2. A theoretical model of user quality metrics for metasearch systems will be produced, which will fit the large amount of empirical data collected.
3. A digital library search engine capable of accepting explicit ranking algorithms for metasearching will be produced and made available as open source software for other research and practical use, which will be based on the theoretical model.

Impacts of Project Outcomes

This research will potentially impact both practical implementations of metasearch systems in research libraries and library consortia, as well as other digital library research efforts concerned with the questions of how to best design metasearch systems. This research will undoubtedly uncover unexpected results that will require subsequent studies to fully comprehend, but it is essential to make a beginning, and to start to develop understanding and solutions. There is definitely potential for this research to have far-reaching impacts that will benefit multiple institutions and constituencies, since virtually all libraries today are confronted with the need for effective metasearching tools; this research will inform the implementation of such systems generally.

SECTION 2: ADAPTABILITY

This research will produce basic findings concerning the expectations and criteria for effective ranking of heterogeneous results in metasearch systems, at least for the broad categories of users that are the target of the study. These results should be generically understandable and capable of being implemented in a wide variety of metasearch software systems, both commercially developed and open source. Targets include WWW searching, library catalogs, OAI-based systems, and organizational portals (for museums, libraries, corporations, and academic/scholarly institutions).

By using a combination of quantitative and qualitative methodologies described below in the design section, the project will attempt to gain perspective on the question of user quality metrics through several approaches. The quantitative study of this aspect of user expectations will provide specific measurements of which factors lead to high quality assessments by users in heterogeneous retrieval sets, while the qualitative analysis aims at understanding user perceptions through cognitive narratives. This combination of research results should provide good insights into this issue for metasearch systems.

SECTION 3: DESIGN

The project will undertake two kinds of investigation into user quality assessments of heterogeneous source metasearch retrieval: 1) quantitative assessment of user quality metrics using an empirical methodology recently employed in previous research, and 2) qualitative assessment through focus groups of the cognitive narratives that users report concerning this issue.

Full details on the sequence of activities are provided in the Schedule of Work Details section of this proposal. All activities will be coordinated by the principal investigators and the Head of Digital Library Research at Emory University.

Experimental Hypothesis

The basic hypothesis to be tested is that end-users of digital library metasearch services think in terms of implicit quality metrics that can be respected in retrieval ranking, improving results over the current text-query relevance and even hub-authority based methods. If effective metasearch systems are to be informed by empirical studies of actual users, this hypothesis must be tested.

Quantitative Methodology

In a recent research study concerning the effectiveness of automatically structured queries [Goncalves, 2004], an innovative technique for analyzing a discrete set of variant search systems was developed. The paper reporting this research is included as Appendix D of this proposal for reference. The experimental methodology used in this previous work will be deployed again in this research project. For our purposes, this methodology involves presenting users querying an experimental metasearch engine with a randomized merged result set of the top results from several different subsystems emphasizing separate (combinations of) information characteristics. Users then are asked to identify the highest quality items from the merged result set, instead of ranking a separate results set for each variant. This avoids bias based on order as well as learning effects. User judgments are recorded and then statistically analyzed to identify the algorithms, factors, and weightings for integrating quality metrics that are statistically preferred by users. Identification of the highest ranked variants produces a good assessment of which quality metrics and which search algorithms users prefer for searching heterogeneous resources. Virginia Tech will take the lead in conducting this quantitative investigation. Statistically relevant numbers of subject responses will be acquired in the course of study.

Qualitative Methodology

Focus groups have been repeatedly used by the MetaScholar Initiative to gain cognitive narratives from users of their expectations of search systems. Previously conducted MetaScholar focus groups did not highlight the question of user expectations of quality ranking of metasearch systems because the issue was not yet clearly identified as problematic. Focus groups in this research project will examine this question directly, engaging users in “think-aloud” exercises and discussions around test systems. The results of the quantitative investigations will be considered in connection with the findings of the focus groups to develop a model of what constitutes user expectations for high quality results in metasearch systems.

User Categories for Study

The focus of this research will be on academic community members, although this focus can be expanded if time permits and a useful population of non-academic subjects can be included. Users will be divided into a two dimensional matrix of categories for purposes of this study: subject expertise and metadisciplinary orientation. Specifically, at least four groups will be studied: 1) graduate students in the humanities, 2) faculty members in the humanities, 3) graduate students in the sciences, and 4) faculty in the sciences.

Heterogeneous Sources of Information for MetaSearching

For humanities subjects, the extended AmericanSouth.Org collection will be used. For scientific subjects, the CITIDEL collection will be used. Both collections are heterogeneous union catalogs based on many disparate source digital libraries. In parallel, using term project groups in senior and graduate level classes at Virginia Tech, we will collect and analyze data related to online library catalogs (OPACs, WebCats) as well as web search engines (e.g., in collaboration with other ongoing research efforts with both academic and commercial groups, e.g. collaborative work that Virginia Tech interns are undertaking with the research division of the Microsoft corporation).

SECTION 4: MANAGEMENT PLAN

Emory University and Virginia Tech are uniquely suited to undertake this project because of a variety of institutional and collaborative factors. A combination of relevant collaborative experience in similar projects and institutional infrastructures provide the capability to successfully carry out the proposed project activities. Full detail on project activity sequencing are provided in the section on Schedule of Work Details.

Institutional Collaborative Relationship: Emory University and Virginia Tech are well prepared to collaboratively embark on this research project, having undertaken several other grant-funded digital library research projects in recent years. Project leaders Halbert and Fox worked closely together on two of the seven projects funded by the Andrew W. Mellon Foundation to advance understanding and usage of the Open Archives Initiative Protocol for Metadata Harvesting [Waters, 2001]. These projects were successfully completed [Halbert, 2003b], and formed the basis of the ongoing MetaScholar Initiative, and laid the groundwork for the OCKHAM project (led by Emory and including VT as partner), sponsored by the National Science Foundation and the Digital Library Federation [Greenstein, 2002], which aims at development of interoperable digital library frameworks. Together with Aaron Krowne, the Emory and Virginia Tech research teams have years of collaborative experience in completing projects successfully together.

Supporting Institutional Infrastructure: Emory University and Virginia Tech have collectively managed millions of dollars of digital library research projects in recent years, and have a strong financial support infrastructure for managing projects. Concerning experimental facilities, Virginia Tech will involve the Digital Library Research Laboratory and the Center for Human-Computer Interaction, with facilities for engaging in usability testing and other computer system evaluation endeavors. Emory University Libraries possess more modest testing facilities, but have routinely hosted focus groups, usability studies, and other kinds of user study groups in recent years. The personnel, facilities, support infrastructure, and experience of the two institutions together provide a pool of resources that can be drawn upon for ad hoc purposes in support of this research project.

SECTION 5: BUDGET

The project budget primarily includes funding for staff and research study expenses, as these are the two critical factors that will enable this project to be successful. Additional information concerning the budget is provided in the budget justification section of this proposal.

Staff: The largest part of the budget is for personnel lines, including research assistants to conduct the experimental studies as well as programming staff to implement test systems. Experience gained in previous similar projects forms the basis for estimating staff time to organize the experiments, program the test systems, and project oversight. Salaries are calculated using typical figures from current project staff and research assistants.

Other Costs: Other project expenses include travel (for attending planning meetings and to present findings), project materials, and minor equipment

SECTION 6: CONTRIBUTIONS

Emory will contribute substantially to the direct costs of this project, in the form of funding for Aaron Krowne's salary. As Head of Digital Library Research at Emory, Krowne will focus all his efforts on the research aspects of this project during the first project year, designing the test systems and experiments that will be used to assess and gather the user data.

SECTION 7: PERSONNEL

Martin Halbert has been Director for Library Systems at the Emory University General Libraries since 1996. He is currently executive director and principal investigator for the projects of the MetaScholar Initiative, an ongoing program for digital library research based at Emory University and focused on applications of metadata harvesting to scholarly communication. He has spoken on the topic of metadata harvesting services at a number of national and international conferences. As PI for this project, he will devote significant time to supervision and coordination of project activities.

Dr. Edward A. Fox holds a Ph.D. and M.S. in Computer Science from Cornell University, and a B.S. from M.I.T. Since 1983 he has been at Virginia Polytechnic Institute and State University (VPI&SU or Virginia Tech), where he serves as Professor of Computer Science. He directs the Internet Technology Innovation Center at Virginia Tech, Digital Library Research Laboratory, Networked Digital Library of Theses and Dissertations, and Computing and Information Technology Interactive Digital Educational Library (CITIDEL). He has been (co)PI on over 80 research and development projects. In addition to his courses at Virginia Tech, Dr. Fox has taught over 50 tutorials in more than 18 countries. He has given more than 35 keynote, banquet, international, invited, and distinguished-speaker presentations, over 75 refereed conference/workshop papers, and over 250 additional presentations. Dr. Fox will supervise and coordinate the research activities of this project at Virginia Tech.

Aaron Krowne holds a M.S. In Computer Science and B.S. degrees in Mathematics and Computer Science from Virginia tech. He is currently Head of Digital Library Research at Emory University, reporting to Martin Halbert. He has previously worked extensively under Dr. Fox on CITIDEL and other digital library research systems. He will supervise the research activities of the project at Emory University.

The Virginia Tech Research Assistant will be a graduate student working in the field of digital library research. The individual recruited will be knowledgeable concerning state of the art digital library research issues and the conduct of user studies.

The Emory Research Assistant will be a graduate student in a field of the social sciences. The individual recruited will be knowledgeable concerning focus group practices and conduct of user studies.

The Emory Programmer will be a professional experienced in software programming. The individual recruited will likely be one of the programmers currently employed in other MetaScholar Initiative projects, and therefore will have pre-existing expertise in programming digital library applications

.

SECTION 8: PROJECT EVALUATION

The three outcomes of the project will be evaluated using the following completion and success criteria.

Completion Criteria (The following criteria must be accomplished to *complete* project work:

1. All user studies described in the design section must be completed, and data made available for free network download on a project website.

2. A systematic analysis of the data from the user studies must be completed that examines the patterns of the results, and this analysis must be submitted in a paper to a major digital library research conference.

3. The prototype digital library search engine capable of accepting explicit ranking formulas for metasearching must be produced and made available as open source software by establishing a site on SourceForge.Net, as well as reporting in other premier access gateways to open source software.

Success Criteria (The following criteria will gauge project *success)*:

1. The data from the user studies must be downloaded by other researchers or practitioners, and some indication must be obtained through feedback (during download, or other venues) that the data was useful in conducting further research or informing decision making concerning metasearch systems.

2. A theoretical model must be produced from the systematic user data analysis. To better serve the needs of users, this model must provide an effective context and framework for understanding implicit user expectations and preferences concerning metasearching, as well as mechanisms for those designing and implementing metasearch services.

3. The prototype digital library search engine software produced must be downloaded and used by one or more other groups researching, designing or implementing metasearch services.

SECTION 9: DISSEMINATION

The completion and success criteria for the project include publication of results in conferences and websites. Some of the specific conferences that are being considered for publication of results include JCDL, ECDL, SIGIR, CIKM, or ICADL. Presentations of results will also be made at meetings of the Digital Library Federation, the Coalition for Networked Information, and the American Library Association.

As mentioned, a SourceForge.Net site will be established for the software, and a project website will make both the software and user study datasets available for free download.

SECTION 10: SUSTAINABILITY

The results of the project will be incorporated into the ongoing work of both the MetaScholar Initiative and the CITIDEL service. This research is timely, and should be of great immediate practical value to many groups investigating metasearching services. The results may become widely utilized by many researchers and practitioners very quickly.

SCHEDULE OF WORK

			2004 Q4			2005 Q1			2005 Q2			2005 Q3			2005 Q4			2006 Q1			2006 Q2			2006 Q3		
		ACTIVITIES	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Phase 1	1	Administrative preparation	🟦																							
	2	Software design		🟦	🟦	🟦	🟦	🟦	🟦	🟦																
	3	Development of user studies research design		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦															
	4	Recruiting for programming and research assistants					🟦	🟦	🟦	🟦	🟦															
	5	Milestone: Software Design and Staff Hires									◇															
Phase 2	6	CVS Code Repository & Other Setup										🟩														
	7	Initial software version programmed										🟩	🟩	🟩												
	8	Tests of initial software using mock user studies											🟩	🟩												
	9	Software revisions and operation												🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩				
	10	Milestone: Completion of Initial Version of Software												◇												
Phase 3	11	Preparation for user studies												🟨	🟨											
	12	Quantitative user studies													🟨	🟨	🟨▪	🟨	🟨	🟨▪						
	13	Qualitative focus group user studies													🟨	🟨	🟨▪	🟨	🟨	🟨▪						
	14	Collation and analysis of data by research assts																🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	
	15	Milestone: Completion of User Studies																				◇				
Phase 4	16	Analysis of data by project staff																				🟥	🟥	🟥▪		
	17	Theoretical modeling																					🟥	🟥▪		
	18	Dissemination of results																					🟥	🟥▪	🟥	🟥▪
	19	Preparation of final report																							🟥	🟥▪
	20	Milestone: Completion of Analysis & Final Report																								◇
		Project Months	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

SCHEDULE LEGEND:

Planning & Preparation =	🟦
Programmer work activity =	🟩
Research Assistant work activity =	🟨
Analysis =	🟥
Milestone =	◇
Presentation opportunity =	▪

Possible presentations:

- 2005 Fall DLF & CNI Meetings
- 2006 Spring DLF and CNI Meetings
- 2006 JCDL Meeting
- 2006 ALA Annual
- 2006 Fall ECDL Meeting
- 2006 ALA LITA Conferenc

SCHEDULE OF WORK DETAIL

PHASE 1: PLANNING AND PREPARATION

1. **Administrative Preparation (October 2004):** Upon notification from IMLS that the grant has been awarded, several administrative steps will be taken. A fund code for tracking all project expenditures will be set up, and the award funding will be deposited into it. Project partners will be notified and initial planning discussions will be scheduled. Deadlines for conferences will be identified for project results dissemination, and proposals for such presentations will be drafted.

2. **Software design (November 2004 – May 2005):** This activity will be performed primarily by Aaron Krowne. Detailed specifications for the software will be documented in order to clearly set out the needs for the search engine. A evaluation process will be undertaken to select an open source search engine for modification and enhancement (a strong possibility is the ESSEX software developed by Krowne, but the evaluation will remain open to other possibilities). Some initial coding of the software may take place during this activity to check for assumptions about the difficulty of producing the software desired.

3. **Development of user studies research design (November 2004 – June 2005):** A detailed plan for the various user studies will be prepared during this activity by Krowne, in collaboration with the principal investigators. Processes for recruiting subjects will be identified, as well as calendar evaluation.

4. **Recruiting for programming and research assistants (February 2005 – June 2005):** The Emory-based programming position will be recruited in this period, as well as the graduate research assistants for both Emory and Virginia Tech.

5. **Milestone: Software Design and Staff Hires (June 2005):** By the end of this month the full software design should be complete, and all project positions should be recruited.

PHASE 2: PROGRAMMING AND DEVELOPMENT

6. **CVS Code Repository & Other Setup (July 2005):** Initial tasks of the programmer will be oriented to preparation for the software development. A code versioning system (CVS) repository will be set up to track versions of the software modules as they are developed and revised. The software design specifications will be discussed in detail with Krowne, and hardware familiarization will take place on the servers to be used.

7. **Initial software version programmed (July – September 2005):** The initial version of the software will be produced in this period. This version will be

capable of all the basic anticipated software needs, but it is inevitable that many requirements will emerge in the course of the actual testing that will require additional modifications over the following months. This activity will also include importing the harvested information from the AmericanSouth and CITADEL portals into the system.

8. **Tests of initial software using mock user studies (August – September 2005):** The initial software will be tested in mock user studies as preparation for the upcoming actual tests. Load testing will take place to assess system performance with full databases. This process will hopefully provide enough feedback for the revisions before the user studies begin.

9. **Software revisions and operation (September 2005 – May 2006):** The search software will be revised following the findings from the test process and the user studies. The programmer will prepare and maintain the test search system during the user studies, coordinating with the research assistants on the conduct of the studies. The establishment of the SourceForge site will also take place during this work activity.

10. **Milestone: Completion of Initial Version of Software (September 2005):** The functioning digital library search engine capable of accepting explicit ranking formulas metasearching will be the major milestone for this phase.

PHASE 3: USER STUDIES

11. **Preparation for user studies (September 2005 – October 2005):** The research assistants at Emory University and Virginia Tech will make preparations for conducting the user studies during this period, familiarizing themselves with the research design plan and project calendar. Arrangements for the studies will be made, including room reservations and initial recruiting of subjects.

12. **Quantitative user studies (October 2005 – March 2006):** Virginia Tech will conduct the quantitative studies during these months. Emory may also participate in the latter half of this period if time permits and the facilities can be arranged.

13. **Qualitative focus group user studies** (**October 2005 – March 2006):** Emory University will host the focus group sessions during this period.

14. **Collation and analysis of data by research assistants (January 2006 – August 2006):** The research assistants will tabulate and document the data collected from the user studies in this period. The Emory research assistant will conclude their work by May 2006, on the assumption that the results of the focus groups will be somewhat easier to summarize. The datasets from the studies will be mounted on the project website by the end of this period. Some analysis by the research assistants will take place in this period as time permits.

15. **Milestone: Completion of User Studies (March 2006):** The completion of all quantitative studies and focus groups will be the major milestone of this phase.

PHASE 4: ANALYSIS AND DISSEMINATION

16. **Analysis of data by project staff (May 2006 – July 2006):** Once the data from the focus groups have been collated, the principal investigators and Krowne will conduct a systematic analysis, looking for patterns using statistical and other tests.

17. **Theoretical modeling (June 2006 – July 2006):** Once an understanding is gained of the patterns (if any) in the user studies data, theoretical modeling will take place.

18. **Dissemination of results (June 2006 – September 2006):** One or more papers will be submitted for publication during this period.

19. **Preparation of final report (August 2006 – September 2006):** The final project report will be prepared jointly by the principal investigators.

20. **Milestone: Completion of Analysis & Final Report (September 2006):** The project will conclude with the completion of research papers reporting findings and the project report.

APPENDICES

Appendix A: Bibliography

Baeza-Yates 1999 Baeza-Yates, Ricardo, and Berthier Ribiero-Neto. Modern Information Retrieval. ACM Press: 1999.

Benatallah 2002 Benatallah, B., M. Dumas, M-C. Fauvet, F. Rabhi. “Towards Patterns of Web Services Composition”. In F.A. Rabhi and S. Gorlatch, eds., *Patterns and Skeletons for Parallel and Distributed Computing*. Springer: 2002.

Borgman 1996 Borgman, Christine L. “Social Aspects of Digital Libraries (working session)”. *Digital Libraries.* 1996: 170. http://is.gseis.ucla.edu/research/dl/UCLA_DL_Report.html

Botafogo 1992 Botafoga, Rodrigo A., Ehud Rivlin, and Ben Shneiderman. “Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics”. *TOIS 10*(2).1992: 142-180

Breese 1998 Breese, John S., David Heckerman, and Carl Myers Kadie. “Empirical Analysis of Predictive Algorithms for Collaborative Filtering”.*UAI*. 1998: 43-52

Brin 1998 Brin, Sergey, Lawrence Page. “The Anatomy of a Large-Scale Hypertextual Web Search Engine”. *WWW7 / Computer Networks 30 (1-7)*. 1998: 107-117

Burgess 2002 Burgess, M., W. A. Gray and N. Fiddian. “Establishing a Taxonomy of Quality for use in Information Filtering”. *BNCOD*. 2002.

Burgess 2003 Burgess, M., W.A. Gray, and N.J. Fiddian. “A Flexible Quality Framework for Use within Information Retrieval”. International Conference on Information Quality. 2003.

Dhyani 2002 Dhyani, Devanshu, Wee Keong Ng, and Sourav S. Bhowmick. “A Survey of Web Metrics”. *ACM Computing Surveys 34(4)*. 2002: 469-503

Fox 1992 Fox, E., S. Betrabet, et al. Extended Boolean Models. In Frakes, W. B. and R. Baeza-Yates, eds., Information Retrieval: Data Structures & Algorithms.. Englewood Cliffs, NJ, Prentice-Hall. 1992: 393-418.

Fox 2001 Fox, Edward A. “Modeling and Building Personalized Digital Libraries with PIPE and 5SL”. *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*. 2001

Frankes 1992 Frakes, W. B. and R. Baeza-Yates. “Information Retrieval: Data Structures & Algorithms”. Englewood Cliffs, NJ, Prentice Hall. 1992.

Garfield 1972 Garfield, E. “Citation Analysis as a Tool in Journal Evaluation”. *Science*. 1972: 178, 471-479

Gonçalves-Dis Gonçalves, M. “Digital Libraries: Formal Theory, Language, Design, Generation, Quality, and Evaluation”. Dissertation chapter draft. 2004.

Gonçalves-TOIS Gonçalves, Marcos André, Edward. A. Fox, Layne T. Watson, Neill A. Kipp. “Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries”. In press to appear in April 2004 issue of *ACM Transactions on Information Systems*.

Goncalves 2004 Goncalves, Marcos, Edward Fox, and Aaron Krowne, et. al. “The Effectiveness of Automatically Structured Queries in Digital Libraries.” Submitted JCDL, 2004.

Gravano 1997 Gravano, L. and H. Garca-Molina. “Merging ranks from heterogeneous internet sources”. Stanford. Stanford. 1997

Greenstein 2002 Greenstein, Daniel, and Martin Halbert. ” Frameworks and Forums for Evolving Digital Library Architectures.” Digital Library Federation, 2002. URL: http://www.diglib.org/architectures/ockham.htm

Grossman 1998 Grossman, D. and O. Frieder. Information Retrieval: Algorithms and Heuristics. Boston, Kluwer Academic Publishers. 1998

Halbert, 2002a Halbert, Martin. “Metadata Gardening: Metadata Aggregation Networks and the MetaScholar Initiative.” Access 2002 Conference, Windsor, Ontario, Canada: Oct 22, 2002. <http://zeus.uwindsor.ca/library/leddy/access2002/martin.ppt>

Halbert, 2002b Halbert, Martin. “The Mellon Metadata Harvesting Initiative: Major Findings from Participating Projects.” DLF Fall 2002 Forum: Seattle, Washington, Nov. 5, 2002. <http://www.diglib.org/forums/fall2002/mellon%20metascholar.htm>

Halbert, 2003a Halbert, Martin. “The MetaScholar Initiative: AmericanSouth.Org and MetaArchive.Org.” Library Hi-Tech 21.2 (2003)

Hansen 1983 Hansen, James V. “Audit Considerations in Distributed Processing Systems”. *CACM 26(8)*. 1983: 562-569

Harman 1992 Harman, D., E. Fox, et al. In Frakes, W. B. and R. Baeza-Yates, eds., Inverted Files. Information Retrieval: Data Structures & Algorithms. Englewood Cliffs, NJ, Prentice-Hall.1992: 28-43.

Ipeirotis 2002 Ipeirotis, Panagiotis G., Luis Gravano. “Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection”. *VLDB*. 2002: 394-405.

Joshi 2000 Joshi, A., R. Krishnapuram. “On Mining Web Access Logs”. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 2000: 63—69 http://citeseer.nj.nec.com/joshi00mining.htm

Kaplan 2000 Kaplan, N. R. and M. L. Nelson. "Determining the Publication Impact of a Digital Library." Journal of the American Society of Information Science 51(4). 2000: 324-339.

Kessler 1963 Kessler, M M. “Bibliographic coupling between scientific papers”. *American Documentation. 1963:* 14:10-25

Kleinberg 1998 Kleinberg, Jon M. “Authoritative Sources in a Hyperlinked Environment”. *SODA.* 1998: 668-677

Liu 2000 Liu, Y.-H., P. Dantzig, et al. "Visualizing Document Classification: A Search Aid for the Digital Library." Journal of the American Society for Information Science 51(3). 2000 216-227.

Lynch 2001 Lynch, Clifford A. “Metadata harvesting and the Open Archives Initiative.” ARL Bimonthly Report Number 217 (Aug 2001). <http://www.arl.org/newsltr/217/mhp.html>

Motro 1998 Motro, Amihai., Igor Rakov. “Estimating the Quality of Databases”. Flexible Query Answering Systems, Third International Conference, FQAS'98. Roskilde, Denmark. May 13-15, 1998. *Lecture Notes in Computer Science* 1495 Springer 1998, ISBN 3-540-65082-2: 298-307

Nowell 1996 Nowell, L., D. Hix, et al. “Visualizing Search Results: Some Alternatives to Query-Document Similarity”. ACM SIGIR '96. Zurich, Switzerland. 1996: 67-75.

Nowell 1993 Nowell, L. and D. Hix. “Visualizing search results: User Interface Development for the Project Envision Database of Computer Science Literature”. Advances in Human Factors/Ergonomics. HCI International 5th International Conference on Human Computer Interaction, Elsevier. 19B. 1993: 56-61.

Pipino 2002 Pipino, Leo, Yang W. Lee, Richard Y. Wang. “Data Quality Assessment”. *CACM 45(4)*. 2002: 211-218

Ramakris 1992 Gonçalves, Marcos André, Ali A. Zafer, Naren Ramakrishnan, Redman, Thomas C. “Data Quality: Management and Technology”. New York: Bantam Doubleday Dell. 1992.

Rittberger 2001 Rittberger, M. “Quality Measuring With Respect to Electronic Information Markets and Particularly Online Databases. New York: Marcel Dekker, 31. 2001: 274-295

Sandhu 1999 Sandhu, Fu Y. and Shih, M.-Y. “Clustering of Web Users Based on Access Patterns”. KDD Workshop on Web Mining, San Diego, CA. Springer-Verlag. 1999 <http://citeseer.nj.nec.com/fu99clustering.html>

Sarasevic 2000 Sarasevic, T. “Digital Library Evaluation: Toward Evolution of Concepts”. Library Trends: Evaluation of Digital Libraries, 49, (2). 2000: 350-369

Suleman 2000 Suleman, H., E. A. Fox, et al. “Building Quality into a Digital Library”. Fifth ACM Conference on Digital Libraries: DL '00. June 2-7, 2000. San Antonio, TX. New York: ACM Press. 2000

Wand 1996 Wand, Yair, and Richard Wang. “Data Quality Dimensions in Ontological Foundations.” Communications of the ACM 39.11. 1996: 86-95.

Wand 2002 Parsons, Jeffery, Yair Wand. “Property-Based Semantic Reconciliation of Heterogeneous Information Sources”. *ER*. 2002: 351-364

Wand 1996 Wand, Yair, Richard Y. Wang. “Anchoring Data Quality Dimensions in Ontological Foundations”. *CACM 39(11)*. 1996: 86-95

Wang 1995 Wang, Richard Y. Veda C. Storey, Christopher P. Firth. “A Framework for Analysis of Data Quality Research”. TKDE 7(4). 1995: 623-640

Wang 2000 Wang, L. and E. A. Fox. “Crawling on the World Wide Web”. Blacksburg, VA, Virginia Tech Department of Computer Science. 2000

Waters 2001 Waters, Donald J. “The Metadata Harvesting Initiative of the Mellon Foundation.” ARL Bimonthly Report Number 217. Aug 2001. <http://www.arl.org/newsltr/217/waters.html>

Witten 2002 Witten, Ian, and David Bainbridge. How to Build a Digital Library. San Francisco, Morgan Kaufmann: 2002.

Appendix B: Literature Review Discussion

Many different aspects of quality in digital libraries have been touched upon in previous work. Quality of navigation services in terms of connectivity, compactness, stratum, depth, and imbalance was discussed in [Botafogo, 1992]. Quality as precision and recall is discussed in [BYRN, 1999]. Quality through recommendation and personalization were the focus of [Breese, Ramakrishan]. Hansen discusses quality as service reliability in [Hansen, 1983]. In [Benatallah, 2002], quality through DL service composability was proposed. [Motro, 1998] and [Ipeirotis, 2002] discuss probing multiple repositories to determine quality of entire repositories. The relationship of quality dimensions to phases of the life cycle of information in digital libraries was discussed in [BorgmanCycle]. Gonçalves discusses quality as conformance of metadata to a standard in [Gonçalves-TOIS]. Kleinberg views quality as mutual reinforcement between items in [Kleinberg, 1998]. Quality as metadata accuracy was discussed in [Redman, 1992].

PageRank, a ranking method based on "reputation" derived from document linkage and used in the search engine Google is discussed in [Brin, 1998]. Co-citation as a relatedness/quality metric is discussed in [Garfield]. Bibliographic coupling is discussed in [Kessler]. Clustering based on access patterns is discussed in [Sandhu], using query log analysis methods as exemplified by [Joshi]. A number of dimensions of quality are identified in [Rittberger] in the context of e-commerce; many of them are quite applicable to digital libraries.

Perhaps closest to our proposed work, [Burgess, 2002] and [Burgess, 2003] discuss a framework for quality use within information retrieval, focusing on models of quality and their representation in XML.

Marcos Gonçalves considers quality in relation to digital libraries in [GonçalvesDis]. In particular, he points out the need to model DL quality in connection with the 5S (societies, scenarios, spaces, structures, streams) framework. With regards to pre-existing work, he writes:

> Regarding quality issues, the Infometrics and Bibliometrics subfields of Library Science utilize quantitative analysis and statistics to describe patterns of publication within a given field or body of literature. As previously argued, the digital nature of DLs brings its own challenges, quality dimensions, and metrics to the table. In computer science, much of the related work has considered the issue of data quality within the database community (e.g., [Redman, 1992, Wand, 1996, Motro, 1998, Pipino, 2002, Wand, 2002]). A comprehensive survey of data quality research [Wang, 1995] has determined that most of the work in data quality has been done in: 1) analysis and design of the data quality aspects of data products; 2) design of data manufacturing systems that incorporate data quality aspects; and 3) definition of data quality dimensions and the measurement of their values. The work proposed here touches some of these aspects but for the digital libraries field, for which such a type of study, mainly one that is based on a formal theoretical foundation, is necessary but has been missing. Finally, Dhyani et al. [Dhyani, 2002] present an extensive survey of Web metrics. While many of the suggested Web metrics can be used or adapted to the DL context (and have been), we have seen that DLs differ from the Web in many ways and therefore a specific study of DL quality dimensions and metrics is necessary.

He also adds:

> DL quality and evaluation is a very underrepresented research area in the digital library literature. Saracevic [Saracevic, 2000] was one of the first to consider the problem. He argues that any evaluation has to consider a number of issues such as the context of evaluation, the criteria, the measures, and the methodology. However, in his analysis, it is concluded that there are no clear agreements regarding the elements of criteria, measures, and methodologies for DL evaluation. As a first attempt to fill some gaps in this area, Fuhr et al. [Fuhr, 2001] propose a description scheme for DLs based on four dimensions: data/collection, system/technology, users, and usage.

We wholeheartedly agree with Gonçalves' appraisal of the situation, though we also would point out that the recent efforts by Burgess et al. [Burgess, 2002, Burgess 2003] seem to represent encouraging steps towards the goal of formalization of quality dimensions in information retrieval.

This brief survey shows that quality is a multifacted concept, even in the special case of digital libraries. For our work, we choose to focus on two aspects: quality of resources and quality of digital library services. We hope to engender the latter by formal and explicit handling of the former, using the 5S framework as Gonçalves has proposed.

Even where research findings in the aforementioned works have been translated into usable systems, these systems were and remain to this day handlers of *special cases* of quality, where only one or a few quality metrics or dimensions were selected. We seek to build systems that will handle resource quality in the *general* sense, thus remaining extensible to future notions of quality.

In addition, no study seems to have been undertaken to discover which metrics of quality are *expected* by users of digital libraries, both casual and expert. Without such data, it is entirely possible that many IR quality efforts have started from the wrong premises. For instance, initial signs seem to indicate that settling on a single dimension of quality to use in a search system, as has been done so far, is completely counter-intuitive to users. Without drawing upon many dimensions of quality, search engine ranks are too distant from the desired “best resource” semantics.

Appendix C: Overview of the MetaScholar Initiative

Summary

The MetaScholar Initiative is a collaborative endeavor to explore the feasibility and utility of scholarly portal services developed in conjunction with OAI metadata harvesting technologies. The MetaScholar Initiative is comprised of several projects, including the *MetaCombine, MetaArchive,* and *AmericanSouth* projects, all funded by grants from the Andrew W. Mellon Foundation totaling in excess of $1M, as well as a recently awarded IMLS project, entitled *Music of Social Change: Library-Museum Collaboration through Open Archives Metadata*. These projects have created metadata aggregation networks connecting some 24 libraries, archives, museums, and electronic text centers. Each network has an associated portal being created under the guidance of teams composed of scholars, librarians, archivists, and technologists. The MetaScholar Initiative is studying issues such as metadata normalization, alternative forms of scholarly communication through portals, and the process of facilitating smaller archival institutions in providing better access to their collections through the OAI-PMH. The MetaScholar Initiative is based at Emory University in Atlanta, Georgia. The following narratives briefly describe the first two projects of the Initiative.

MetaArchive Project

A consortium of educational libraries, archives, and museums led by Emory University is undertaking a demonstration project and feasibility study for a cross-institutional scholarly portal service providing public search functionality and subject organization for archival metadata aggregated using the OAI-PMH. The project arose from the belief that OAI-based services for researchers must do more than simply aggregate metadata. To be effective tools for research, there must be framing organization, contextual materials, and other sorts of information that add value to the basic functions of metadata aggregation and search. Further, a great advantage of such services is that they might fruitfully be focused on specific subject domains in order to concentrate and leverage attention and knowledge that subject specialists contribute to online gathering spaces. This perspective applies accumulated lessons from decades of online community research to the new opportunities provided by metadata harvesting technologies.

The project set out to actively aggregate metadata by providing partner institutions with direct assistance in the form of data conversion expertise and programming of OAI provider systems. The project would seek to convert and import existing metadata, in the form of finding aids, catalog records, or other machine-readable forms. Typical partner institutions are archives of four-year liberal arts colleges, which frequently have only one archivist and lack technical staff or infrastructure for sharing metadata via the OAI-PMH. This category of smaller archive was a major focus because there are a very large number of such repositories that collectively hold information of great interest for scholarly research, but for which there are inadequate mechanisms for cross-institutional discovery of resources, as evidenced by the case studies described above. For perspective on the issues of catalyzing metadata provision through the OAI-PMH in other sorts of institutions, the project also sought to work with a small number of larger archives in research universities and museums. It was anticipated that this activity of metadata aggregation would go on throughout the duration of the project, as some sites would be relatively easy and others would take more time. *MetaArchive partner institutions* include: Southwestern University, the United Methodist Archives, the Atlanta History Center museum, the University of the South, Davidson College, Washington & Lee University, the University of Richmond, and sub-units of Emory University and the University of Georgia.

The *MetaArchive Subject Portal Working Group* (SPWG) has extensively considered the question of how aggregated metadata can be made coherent and useful to scholars. The group is comprised of librarians and archivists at Emory University with doctoral level subject qualifications and extensive experience in the field. The group is studying various design questions concerning scholarly portals and metadata services. Participants include: Linda Matthews (Special Collections Director), Susan Bailey (Bibliographic Gateway Service Division Leader), Raquel Cogell (Research and Information Services Team Leader), Randall Burkett (Ph. D. Religion, Archivist, Special Collections), Steven Ennis (Ph. D. English, Curator of Literary Collections, Special Collections), Joel Herndon (Ph.D. Political Science, Head, Electronic Data Center), Alice Hickcox (Ph.D. Religion, Lewis H. Beck Electronic Text Center), Naomi Nelson (Ph.D. History, Congressional Archivist, Special Collections) Marie Nitschke (Ph. D. History, Reference Librarian), Susan Pinckard, General Libraries Monographic Cataloging Team Leader), and Charles Spornick (Ph.D. Medieval History, Head, Lewis H. Beck Electronic Text Center).

AmericanSouth Project

The AmericanSouth project seeks to create a definitive scholarly portal for Southern history and culture. The project is the result of an extended planning effort by SOLINET. The project is similar in many ways to the MetaArchive project in that it proposed layering portal services on top of a central metadata harvester that would aggregate information from cooperating partner libraries. A team of senior researchers (rather than librarians or archivists) serves as the Scholarly Design Team for the project (see below) providing the intellectual organization for this scholarly portal, designing an interactive structure to promote and facilitate research, teaching, and communication.

AmericanSouth seeks to establish OAI provider systems at large research libraries situated around the Southeast. The goal is to create an extensive base of information useful for Southern cultural studies by aggregating metadata from many important archival collections held by ten *AmericanSouth partner institutions:* Auburn University, Emory University, Louisiana State University, the University of Florida, the University of Georgia, the University of Kentucky, the Kentucky Virtual Library, the University of North Carolina at Chapel Hill, the University of Tennessee at Knoxville, and Vanderbilt University. Large research libraries were selected that had significant technical staff and infrastructures capable of installing and maintaining their own OAI provider systems, at least provided that they received a certain level of catalytic assistance by expert OAI consultants. These consultants include computer scientists from Virginia Tech who are participating in the core design work of the OAI protocol.

The design of the AmericanSouth.Org portal is guided by a *Scholarly Design Team* (SDT) of five major scholars of Southern culture and history who have been recruited to think systematically about the issues entailed in such a system. What are the requirements for an authoritative online portal for scholars? What new forms of scholarly discourse are facilitated by an online community built in close proximity to organized metadata concerning primary research materials? How can contextual functions such as annotation, interpretation, and methodological/pedagogical guides be contributed for such metadata aggregation services? In directly recruiting a team of major scholars to work on the project, AmericanSouth.Org is unlike many other such projects, which are primarily driven by technologists. The project team has consistently felt that this scholarly involvement represents a significant strength of the project.

AmericanSouth.Org is fortunate in that from the beginning the project proposal identified Dr. Charles Reagan Wilson as chair of the SDT. Dr. Wilson is a noted scholar of the South and a capable academic leader of collaborative activities. As the primary editor of *The Encyclopedia of Southern Culture* (perhaps the pre-eminent comprehensive guide to scholarship in this area), and director of the Center for the Study of Southern Culture at Ole Miss, Dr. Wilson is well suited academically and professionally for the role he has been called upon to play. His insights and advice have from the start been invaluable, and he has been a true partner in conceptualizing the services that AmericanSouth.Org might provide.

Dr. Allen Tullos of Emory University is an interdisciplinary scholar engaged in many research areas, including American popular culture, the South, cultural geography, and the effects of networked information on society. He is noted for developing the *American Routes* public scholarship site (see http://amroutes.cc.emory.edu), a collaborative linkage between an instructional website at Emory and one of the most popular nationally syndicated radio programs.

Dr. Will Thomas of the University of Virginia is one of the developers of the Valley of the Shadow website (http://www.iath.virginia.edu/vshadow2). As one of the pioneers in new forms of online historical scholarship, Dr. Thomas brings to the project years of experience and perspective concerning the possibilities of the new medium.

Dr. Lucinda MacKethan of North Carolina State University was co-editor of the comprehensive *Companion to Southern Literature*, as well as the *Scribbling Women* teaching website (see http://www.scribblingwomen.org). Dr. MacKethan brings to the project a variety of expertise to the project including comparative literature, gender issues in scholarship, and pedagogy of new media.

Dr. Carole Merritt is Director of the Herndon Home (http://www.theherndonhome.org), a national historic landmark memorial site celebrating the heritage of African Americans in the South. Dr. Merritt is a respected scholar of African American family history and culture. She speaks authoritatively to several issues, including public scholarship and African American culture.

Appendix D:

This appendix provides copies of published and forthcoming papers written by project staff in collaboration with others. These papers have informed this proposal in both methodological and theoretical ways.

Link Fusion: A Unified Link Analysis Framework for Multi-Type Interrelated Data Objects

Wensi Xi[1], Benyu Zhang[2], Zheng Chen[2], Yizhou Lu[3], Wei-Ying Ma[2], Edward A. Fox[1]
[1]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24061, U.S.A.
{xwensi, fox}@vt.edu
[2]Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, P.R. China
{i-benyu, zhengc, wyma}@microsoft.com
[3]School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China
luyizhou@pku.edu.cn

ABSTRACT

Web link analysis has proven to be a significant enhancement for quality based web search. Most existing links can be classified into two categories: intra-type links, which represent the relationship of data objects within a homogeneous data type, and inter-type links, which represent the relationship of data objects across different data types. Unfortunately, most link analysis research only considers one type of link. In this paper, we propose a unified link analysis framework, called "link fusion", which considers both the inter- and intra- type link structure among multiple-type inter-related data objects and brings order to objects in each data type at the same time. The PageRank and HITS algorithms are shown to be special cases of our unified link analysis framework. Experiments on an instantiation of the framework that makes use of the user data and web pages extracted from a proxy log show that our proposed algorithm could improve the search effectiveness over the HITS and DirectHit algorithms by 24.6% and 38.2% respectively.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval; G.2.2 [**Discrete Mathematics**]: Graph theory

General Terms

Algorithms, Experimentation

Keywords

Link fusion, Link analysis algorithms, Information retrieval, Data fusion.

1. Introduction

The World Wide Web is estimated to contain 3-5 billion web pages nowadays and is still growing at a rate of 10 million per day. The content of web pages ranges from "Joe Neighbor's" dinner plan to the proceedings of the W3C conference. With such huge volume and great variation in contents, finding useful information effectively from the web becomes a very challenging job. Traditional "keyword based" text search engines cannot provide satisfying results to web queries since: (1) Users tend to submit very short, sometime ambiguous queries and they are reluctant to provide feedback information [3]. (2) The quality of web pages varies greatly [6], and users usually prefer high quality pages over low quality pages in the result set returned by the search engine. (3) A non-trivial number of web queries target at finding a "navigational starting point" [9] or "URL of a known-item" [8] on the web. Thus, web pages containing textually "similar" content to the query may not be relevant at all.

Based on the observations above, researchers tried different approaches to improve the effectiveness of web search engines. One of the representative solutions is re-ranking the top retrieved web pages by their importance [1, 11, 17], which is calculated by analyzing the hyperlinks among web pages. Hyperlink analysis (such as [1, 3-7, 17-19]) has been shown to achieve much better performance than full text search, in production systems.

According to their types, links can be classified into two categories: intra-type links, which represent the relationship of data objects within a homogeneous data space, and inter-type links, which represent the relationship of data objects between heterogeneous data spaces. Most current web link analysis research only analyzes the hyperlinks within web pages, which can be considered as a homogeneous data space. But in the real world, the web pages will often interact with other types of objects, such as users and the queries. In this paper we try to deal with these inter-relationships by expanding the link analysis to combine both inter-type link analysis and intra-type link analysis, and thereby improve web search performance. In Figure 1, we show an example of inter and intra type links by analyzing the relationship of three related data types in the web environment: user, web page, and query.

Users and the queries they submit, plus the web pages they browse, form three homogeneous data spaces. They are correlated when a user submits queries, a user browses web pages, and a query references web pages. The three operations: submit, browse, and reference, involve inter-type links across these data spaces. The hyper-links within web pages, content-based similarity of queries, and social structure of users are intra-type relationships within each space. It is obvious that when analyzing the attributes of web pages, not only the hyper-links between them, but also the users who browse them and the queries that reference them can play important roles.

W3C Conference 2004, May 17-22, 2004, New York, NY USA.
ACM Press

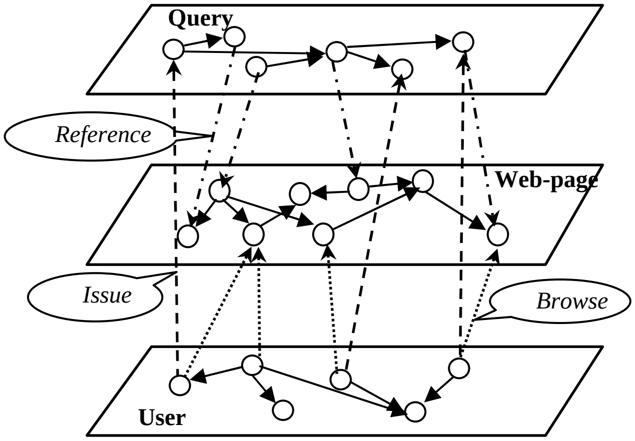

Figure 1: An example of multi-type interrelated data spaces

Most existing web related research fits into the web multi-space model we described in Figure 1. For example, web search [4, 17] uses the web page space and hyperlinks within the space; collaborate filtering [14] uses the document (web page) space, the user space, and the browsing relationship in-between; web query clustering [22] uses the web page space, query space, and reference relationship in-between. Unfortunately, most of these works only consider one type of link/relationship when analyzing the links/relationships of objects, and they can be classified into intra-type link analysis and inter-type link analysis regarding the type of links they use. In intra-type link analysis, the attribute of a data object is directly reinforced by the same attribute of other data objects in the same data space. For example, in Google's PageRank algorithm [4], the "popularity" attributes of web pages are reinforcing each other via the hyper-link structure within them. In inter-type link analysis, the attribute of one type of data objects is reinforced by attributes of data objects from other data spaces. (Examples of inter-type link analysis will be given in Section 3.) Hyperlink analysis reflects the attributes of web pages from the editor's view. The assumption of the hyperlink analysis is that users agree with the editor/author of the web pages in terms of the link structure. It may not work well when a user's perception of a web page differs from that of the authors/editors. Another example of inter-type link analysis, the DirectHit algorithm [11], well captures the web user's view of the web pages from their interactions with the Web. DirectHit utilizes the inter-type links provide by end-users for web search assuming that the more frequently users visit a web page the more important the web page is.

It is natural to ask: Is it possible to combine the process of intra-type link analysis for the same data type and inter-type links across different data types together to improve the process of understanding the organizational relationship of data objects and finding the correct order of data objects regarding different attributes in multiple data types? Intuitively, a simple way is to calculate the data object attributes using inter-type and intra-type link analysis individually, and then combine the results together. However, this solution does not fully utilize the fact that inter- and intra- type links may reinforce the attribute of a data object at the same time. Hence, a unified framework for link analysis is proposed in this paper. The assumption is that the attribute of a data type is influenced not only by the intra-links of its own type but also influenced by the inter-links from other attributes of other different data types. Furthermore, different attributes of different data types can reinforce each other. The problem of leveraging link structures within and across different data types to gain more understanding of the organizational structure and attribute order of objects within each data type can be referred to as the "**Link Fusion**" problem. This name is borrowed from the concept of "**Data Fusion**" in information retrieval where multiple sources of evidences are combined in order to improve the prediction of the relevanc of documents to a query. Experiments on an instantiation of the framework that makes use of users and web pages from a proxy log show that by using our approach, the search precision is improved by 24.6% and 38.2% compared to the traditional HITS [17] and DirectHit [11] algorithms, respectively.

The rest of this paper is organized as follows. In Section 2, we present related work on current state-of-the-art link structure analysis algorithms. In Section 3, we present the proposed unified link analysis framework for multi-type inter-related data objects, which can support HITS and PageRank, as well as the DirectHit algorithm. Then, we show the experimental results in Section 4. Finally, we conclude in Section 5.

2. Related Works

Research on analyzing link structures to better understand the informational organization within data spaces can be traced back to research on "Social Networks" [13]. A good example comes from the telephone bill graph. By searching connected and isolated components, scientists can estimate the diameter of the whole graph and hunt for each complete sub-graph or "clique", to indicate contacts among people. Another interesting example is the famous sociology phrase "six degree of separation", which means that any pair of people on the earth can get acquaint through no more than six intermediaries. Although proving this is still far from complete, some sub-graphs of human society can be explored easily and thoroughly. For instance, members of an enterprise can form an operation graph. By recognizing the functional relationship of each employee, one can learn structural and relative "importance" of each employee within the organization. The problem of link structure of social networks can be reduced to a graph G = (V, E), where set V refers to people, and set E refers to the relationship among people. Katz [16] tried to measure the "importance" of a node in a graph by calculating the in-degree (both direct and indirect) of that node. Hubbell [15] tried to do the same thing by propagating the "importance" weights on the graph so that the weight of each node achieves "equilibrium".

Researchers from the bibliometrics area claimed that scientific citations could be regarded as a special social network, where journals and papers are the nodes and the citation relationships are edges in the graph. Garfield's famous "impact factor" [12] calculates the importance of a journal by counting the citations the journal received (the in-link) within a fixed amount of time. Pinski and Narin [20] claimed that the importance of a journal is recursively defined as the sum of the importance of all journals that cited it. Based on this hypothesis, they designed the following measure of importance. Consider matrix A is the link matrix in the journal space. A_{ij} denotes the fraction of the number of citations from journal *i* to journal *j*. Suppose w_j is the importance value of journal *j*, their calculation can be represented as $w_j = \sum_i A_{ij} w_i$ By iteratively calculating the formula above, it leads to $A^T w = w$, where w is the vector of important weights of journals. It is easy to find out that w is the principle eigenvector of A^T. Following the same rationale, Brin and Page [4] design the PageRank algorithm to calculate the importance of web pages in the Web. In addition to Pinski and Narin's algorithm, PageRank simulates a web surfer's behavior on the web. That is, with probability 1-ε, the surfer randomly picks one of the hyperlinks on the current page and jumps to the page it links to; with probability ε, the user "resets" by jumping to a web page picked uniformly and at random from the collection. This defines a Markov chain on the web pages, with the transition matrix $\varepsilon U + (1-\varepsilon)M$, where U is the transition matrix of uniform transition probabilities ($u_{ij} = 1/n$ for all *i*, *j*). The vector of PageRank scores w is then defined to be the stationary distribution satisfying $(\varepsilon U + (1-\varepsilon)M)^T w = w$. Adding the random surfer model can prevent the "sink node problem" in the PageRank calculation.

Kleinberg [17] claimed that web pages and scientific documents are governed by different principles. Journals have approximately the same purpose, and highly authoritative journals always refer to other authoritative journals. The World Wide Web, however, is heterogeneous, with different pages serving different roles. Authoritative web pages do not necessary link to other authoritative pages, thus Pinski and Narin's hypothesis for scientific literature does not hold in the web. Based on his observations, Kleinberg divides the notion of "importance" of web pages into two related attributes: "Hub" (measured by the "authority" score of other pages that a page links to), and "Authority" (measured by the "hub" score of the pages that link to the page). Different from the PageRank algorithm which calculates the importance of web pages independently from the search query, Kleinberg presented his Hyperlinked-Induced Topic Search (HITS) algorithm as following: (1) Use an ordinary search engine to search the query and form the root set as the starting point; (2) Get the base set by adding pages pointing to or pointed at root pages; (3) Count the authority and hub weights of each page in the base set with an iterative algorithm: for each page, let *a*(*p*) and *h*(*p*) denote its authority attribute weight and hub attribute weight. The two attributes can be calculated as:

$$a(p) = \sum_{q \to p} h(q) \quad and \quad h(p) = \sum_{p \to q} a(q)$$

Let A denote the adjacency matrix of the base set: a_{ij}=1 if page *i* has a link to page *j*, and 0 otherwise. Vectors *a* and *h* correspond to the authority and hub scores of all pages in the base set, hence, a=$A^T h$ and h=Aa. It is easy to show that *a* and *h* are eigenvectors of matrices A^TA and AA^T. The search system [1] developed using the HITS algorithm achieves comparable performance with "Yahoo!", which maintains a manual compilation of net resources. Many researchers have extended the HITS algorithms to improve its efficiency. Chakrabarti et al. [5, 6] used texts that surround hyperlinks in source web pages to help express the content of destination web pages. They also reduce weight factors of hyperlinks from the same domain to avoid a single website dominating the results of HITS. Ng et al. [19] presented randomized HITS and subspace HITS algorithms to enhance the stability of the basic HITS. The former imitates a random walk on web pages and defines the authority/hub weight as a chance of visiting that page in time step *t* (*t* is large enough). The latter uses the first *k* eigenvectors instead of the entire matrix A^TA to count the authority values. Cohn et al. [7] introduced a probabilistic factor into HITS and applied the EM model. All these show that the authority idea has great potential in web applications.

Inter-type links (links that connect different types of data objects) represent relationships of different domains. Researchers also analyzed this kind of link to find out whether it can help improve the link analysis of the data objects within the same data type. For example, DirectHit [11] harnesses the web pages visited by millions of daily Internet searchers to provide more relevant and better-organized search results. Based on the assumption that the most relevant pages of a topic are those most visited, DirectHit's ranking algorithm is used by Lycos, Hotbot, MSN, Infospace, About.com, and roughly 20 other search engines. Miller [18] proposed a modified HITS algorithm, which also utilizes the users' behavior on the web to improve the calculation of hub and authority scores. In his algorithm, the adjacency matrix A is modified and the value of a_{ij} in A is increased whenever a user travels from page *i* to page *j* (information obtained by analyzing web-site access logs). Although Miller uses links from two different spaces (user and web space), he only converted inter-type links (links between users and web-pages) to intra-type links (links within web-pages) to enhance the link analysis for web pages. The users' importance is ignored in this algorithm.

Most recently, Davison [10] analyzed multiple term document relationships by expanding the traditional document-term matrix into a matrix with term-term and doc-doc sub-matrices in the diagonal direction and term-doc and doc-term sub-matrices in the anti-diagonal direction. The term-term sub-matrix represents term relationships (e.g., term similarity), and the doc-doc sub-matrix represents document relationships (e.g., link matrix for web pages). He proposed that the links of the search objects (web-page or terms) in the expanded matrix could be emphasized. With enough emphasis, the principal eigenvector of the extended matrix will have the search object on top with the remaining objects ordered according to their relevance to the search object. Considering that terms and documents each form a different data space, with the doc-term and term-doc matrices representing inter-type links, and the term-term and doc-doc matrices as intra-type links, Davison's proposed research fits our framework very well.

3. The Link Fusion Algorithm

There are similarities among link analysis in social networks, scientific citations, and hyperlink analysis in the web. The data objects in these examples form one or multiple data spaces of different types. Each data space contains one specific attribute of data. Researchers take advantage of the links/relationships either within each data space (intra-type links) or across different data spaces (inter-type links) to calculate the specific attribute of the objects in each of the data spaces. In this Section, we generalize previous link analysis studies and propose a unified link analysis framework to calculate the attributes of data objects within multiple data spaces. We call this unified link analysis framework "Link Fusion algorithm".

Suppose we have *n* different types of objects X_1, X_2… X_n. Each type of data object X_i contains a specific attribute F_i. Data objects within the same type are interrelated with intra-type relationships $R_i \subseteq X_i \times X_i$. Data objects from two different types are related with inter-type relationships $R_{ij} \subseteq X_i \times X_j$ ($i \neq j$). Suppose attributes of different types of data objects are comparable (e.g., similar in nature). We borrow and extend Pinski and Narin's recursive definition of importance [20] and define that the specific attribute of a data object in one data type equals the sum of the attributes of other data objects in the same data space that link to it, plus the sum of other related attributes of data objects in other data spaces and links to it, mathematically as:

$$F_i = F_i R_i + \sum_{j \neq i} F_j R_j \tag{1}$$

For simplicity, we first explain the case that only contains two types of related objects as example to illustrate Eq. (1). We consider two types of objects $X = \{x_1, x_2, \cdots x_m\}$, and $Y = \{y_1, y_2, \cdots y_n\}$ and relationships of R_X, R_Y, R_{XY} and R_{YX} . The adjacency matrices are used to represent the link information. L_X and L_Y stand for the adjacency matrices of link structures within set X and Y, respectively. L_{XY} and L_{YX} stand for the adjacency matrix of links from objects in X to objects in Y and adjacency matrix of links from objects in Y to objects in X respectively. $L_{XY}(i, j) = 1$, if there is a link from node x_i to node y_j , and $L_{XY}(i, j) = 0$ otherwise. Suppose w_x is the attribute vector of objects in X, w_y is the attribute vector of objects in Y, Eq. (1) can be mathematically represented as:

$$\begin{cases} w_y = L_y^T w_y + L_{xy}^T w_x \\ w_x = L_x^T w_x + L_{yx}^T w_y \end{cases} \tag{2}$$

and it can be easily extended into N interrelated data spaces, as shown in Eq. (3)

$$w_M = L_M^T w_M + \sum_{\forall N \neq M} L_{NM}^T w_N \tag{3}$$

There are two issues that need to be considered in Eq. (3):

First, as noted by Bharat and Henzinger [3], mutually reinforcing relationships between objects may give undue weight to objects. Ideally, we would like all the objects to have the same influence on the other objects they connect to. This can be solved by normalizing the binary adjacency matrix in such a way that if an object is connected to *n* other objects in one adjacency matrix, each object it connects to receives 1/*n* of its attribute value. The random surfer model used in PageRank also can be introduced here to simulate random connection, and avoid sink nodes during the computation.

Second, it is too naïve to assume that attributes from different data spaces are equally important, when used to calculate the attribute of data objects. This can be solved by changing Eq. (2) into a weighted sum of attributes. With the consideration of the two issues above, Eq. (3) can be further improved into Eq. (4):

$$\begin{cases} w_M = \alpha_M L_M'^{\,T} w_M + \beta_{NM} \sum\limits_{\forall N \neq M} L_{NM}'^{\,T} w_N \\ where \\ \alpha_M + \sum\limits_{\forall N \neq M} \beta_{NM} = 1; \quad \alpha_M > 0 \quad \beta_{NM} > 0 \quad ; \\ L_M' = \varepsilon U + (1-\varepsilon) L_M ; \quad 0 < \varepsilon < 1 \quad ; \\ L_{NM}' = \delta_N U + (1-\delta_N) L_{NM} ; 0 < \delta_N < 1. \end{cases} \tag{4}$$

In Eq. (4), U is the transition matrix of uniform transition probabilities ($u_{ij}=1/n$ for all *i, j*; where *n* is the total number of objects in data space N). δ and ε are smoothing factors used to used to simulate random relationships in matrices L_M and L_{NM}. L_M and L_{NM} are normalized adjacency matrices.

As with the PageRank and HITS algorithms, the attribute value of objects in our framework can be obtained by iteratively calculating Eq. (4) until the result converges. With the definition of Eq. (4), we actually created a unified square matrix A, as shown in Eq. (5), where n is the total number of all involved objects in different data spaces. The unified matrix A has L'_M on the diagonal direction, and L'_{NM} in other parts of the unified matrix as illustrated below.

$$A = \begin{vmatrix} \alpha_1 L_1' & \beta_{12} L_{12}' & \cdots & \beta_{1n} L_{1n}' \\ \beta_{21} L_{21}' & \alpha_2 L_2' & \cdots & \beta_{2n} L_{2n}' \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n1} L_{n1}' & \beta_{n2} L_{n2}' & \cdots & \alpha_n L_n' \end{vmatrix} \tag{5}$$

Suppose *w* is the attribute vector of all the data objects in different data spaces. The proposed iterative approach is actually transforming the vector *w* using matrix A (e.g., $w = A^T w$). It is relatively easy to find out that when the calculation converges, *w* is the principle eigenvector of matrix A. The formal mathematical proof of the convergence of the calculation can be found in the appendix. Two problems need to be addressed in the construction of the unified matrix A.

Suppose M and N are two heterogeneous data spaces, when a data object in M has no linking relationship to any data objects in N, we set all the elements in the corresponding row of the sub-matrix $L'^{\,T}_{NM}$ to 1/*n*, where n is the total number of objects in data space N. The reason we use random relationship to represent no relationship is to guarantee all the sub-matrix $L'^{\,T}_{NM}$ to be non-zero and to prevent "sink nodes" that may eat up all the weights during the calculation (as suggested by the PageRank algorithm). However, in practice, we can always ignore undesired intra/inter type relationships by setting the corresponding α or β to 0.

In the unified matrix, if $\beta_{MN}>0$, then $\beta_{NM}>0$. This is a necessary condition for the recursive calculation to converge, (as explained in the appendix). However, if the relationship of $L'^{\,T}_{NM}$ is really undesirable for the link analysis, we can always assign a very small positive β_{NM} to reduce the effect of $L'^{\,T}_{NM}$.

By constructing a unified matrix using all the adjacency matrices, we actually construct a unified data space, which contains different types/attributes of data objects. Previous inter-type links are now intra-type links in the unified space, and the "link fusion algorithm" is reduced to link analysis in a single data space.

The proposed framework can be easily used to explain previous works on link analysis.

The PageRank algorithm can be considered as a special case of our unified link analysis framework. In PageRank, there is only one attribute (popularity) of one kind of data object (web pages) being considered. Having α=1 and β=0, (4) reduces to $w = L^T w$ which is the original definition of PageRank algorithm.

The HITS algorithm also can be considered as a special case of the unified link analysis. In the HITS algorithm, two attributes (hub and authority) of the same type of data objects (web pages) are being considered. Hub attributes and authority attributes of the same set of web pages each form a data space; the hyperlinks in-between web pages are now inter-type links that connect the Hub space and Authority space as illustrated in Figure 2.

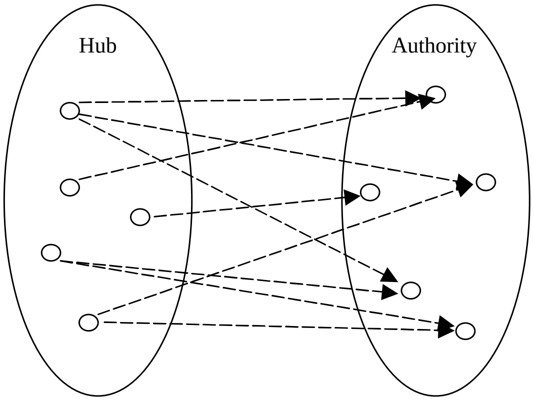

Figure 2. Hub and Authority spaces in HITS algorithm

Since there are not intra-links in each data space, we set α=0 and β=1 and derive the recursive updating equation from Eq. (4): $w_a = L_{ha}{}^T w_h$ and $w_h = L_{ah}{}^T w_a$, where w_a is the authority value vector, w_h is the hub value vector and L_{ha} L_{ah} are adjacency matrices. Considering the normalization of the adjacency matrices and the introducing of smoothing factor ε , this is by definition the Randomized HITS algorithm [19], which is more robust and stable than the traditional HITS algorithm.

4. Experiments

4.1 Experimental Data Set

We use 10 days log from a proxy server at Microsoft to evaluate the effectiveness of our proposed Link Fusion algorithm. The raw proxy logs records user visit information, in which one record corresponds to one HTTP request for a web object from an IP address. In other words, different users from the same IP address are considered as the same user in our experiments. Some heuristic rules (e.g., the words within the hyperlinks, the extension of the filenames, etc.) are applied to filter out the unrelated information, (e.g., ads, images, etc.). Only text pages are reserved in the final dataset, which contains 2,998,821 visit records to 1,773,718 pages by 38,887 users.

4.2 Experimental Approach

Our goal is to improve the end-user's search effectiveness through re-ranking the search results by our proposed Link Fusion algorithm. In order to fit into our framework, we extended the underlying assumption of the HITS algorithm to incorporate the notion of user's "popularity" attribute, and it is defined as below:

- A popular user always look at good hub and good authority pages;
- A good Hub page always points to good Authority pages and is always visited by popular users;
- A good Authority page is always pointed at by good Hub pages and is always visited by popular users too.

The Hub, or Authority attribute of web pages, and the Popularity attribute of users form three different data spaces. These three data spaces are correlated via the hyper links between web pages and user access information from the web proxy log. Their relationships are more clearly illustrated in Figure 3 below.

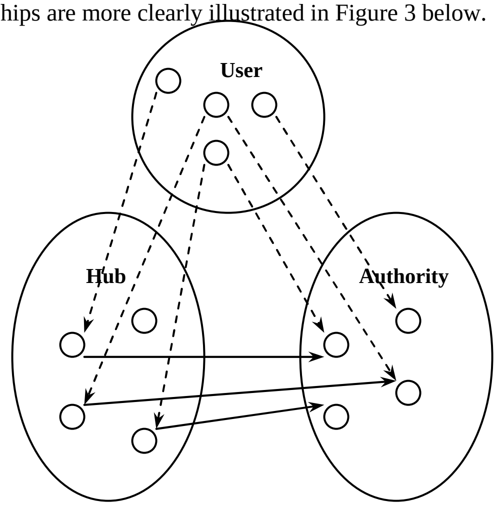

Figure 3. Hub, Authority and User spaces

We find that the three data spaces and the links in-between them fit our Link Fusion algorithm perfectly. We apply the Link Fusion algorithm from Eq. (5) into this case, and derive the unified adjacent matrix as Eq. (6):

$$A = \begin{vmatrix} \alpha_u L_u^{'} & \beta_{uh} L_{uh}^{'} & \beta_{ua} L_{ua}^{'} \\ \beta_{hu} L_{hu}^{'} & \alpha_h L_h^{'} & \beta_{ha} L_{ha}^{'} \\ \beta_{au} L_{au}^{'} & \beta_{ah} L_{ah}^{'} & \alpha_a L_a^{'} \end{vmatrix} \qquad (6)$$

where the sub-scripts a, h and u denote the Authority, Hub, and User space respectively. Since in our case, each data space has no intra-links, we set α_i = 0 (i = a, h, u), and we set all the β equal to 0.5. The initial attribute value of each object is set to 1/n, where n is the total number of objects in the corresponding data space N. Suppose w is the attribute value vector of all the data objects in the three spaces, their final attribute values in w can be obtained by recursively calculating w^{i+1}=$\mathrm{A}^T w^i$ (where i is the iteration number) until converge (e.g., $d = \left\| w^{i+1} - w^i \right\|_1$ is smaller than a thresh-hold value)

After generating the link matrix, we calculate the different attributes of web pages and users and use the "Authority" attribute of web pages to re-rank the search results. The detailed approach is described as follows.

We choose 10 sample queries (shown in Figure 1.) to evaluate the Link Fusion algorithm. Detailed experiment steps for each of the sample queries are:

Step 1: Creating the Hub space and Authority space. The Hub space and Authority space are constructed in a way similar to the HITS algorithm. That is, the query is first sent to a text-based search engine, and the top 200 matching web pages are retained as the root set. Then, the root set is expanded to the base set by its neighborhoods, which are the web pages that either point to or are pointed at by pages in the root set. In this experiment, we set the maximum in-degree of nodes as 50, which is commonly adopted by the previous works [3, 17]. The expanded set of web pages forms the data objects in Hub space and Authority space. Hyperlinks between web pages not on the same web site form the directed links connecting the Hub and Authority space.

Step 2: Creating the User space. After we created the Hub/Authority spaces, we compare the web pages in these spaces with the MSN proxy log data, and extract out the overlapping web pages. The users who browsed these overlapping web pages form the User space, and their browsing activity forms the links from the User space to the Hub/Authority space.

In this experiment we tried to select popular queries to increase the overlapping pages of the Hub/Authority space and the proxy log, so as to increase the number of links between user space and Hub/Authority space. The queries we selected are shown in Table 1.

Table 1. Queries used in experiments

ID	Query	PN	LN	UN
1	search engine	3756	406	9317
2	telephone service	3969	320	20406
3	audi car	2438	220	15369
4	baby care	6050	419	7637
5	windows XP	2288	788	16892
6	computer vision	6116	440	10289
7	notebook computer	3071	299	7810
8	online dictionary	5529	324	8255
9	network security	4762	514	14054
10	daily news	3762	367	8387

In Table 1, PN denotes the total number of pages in the formed Hub/Authority space. LN is the number of pages in the Hub/Authority space that were linked by User space (or the number of links from User to Hub or Authority Space). UN denotes the total number of different users in the User space.

Step 3: Calculation. After creating all three data spaces, we assign an initial weight to each data object, as introduced in Section 4.1. and start the recursive calculation on the different attribute in the data spaces according to Eq. (6) until convergence.

Step 4: Evaluation. Finally, we re-rank the top returned documents according to the Authority value we derived from recursive calculation of w^{i+1}=A$^{T}w^{i}$. Then we use precision at top 10 documents to compare our results with other algorithms.

4.3 Results Evaluation

In this section, we compare the performance of Link Fusion algorithm with that of the text-based retrieval algorithm, HITS algorithm and DirectHit algorithm. More specifically, for each of the queries, the union set of top 10 documents returned from the 4 algorithms are pooled together and rated for relevance by 5 volunteers. The final relevance judgment for each <query, document> pair is decided by majority votes (e.g., the pair is relevant only if more than 3 volunteers voted it as relevance). We then computed precision at top 10 documents (p@10) for each of the four algorithms. This measurement is defined as: $p@10 = r/10$, where r is the number of relevant documents in the top 10 pages returned. The comparison of precision for 4 algorithms is shown in Figure 4. The label "avg" is the average *p@10* across the 10 queries.

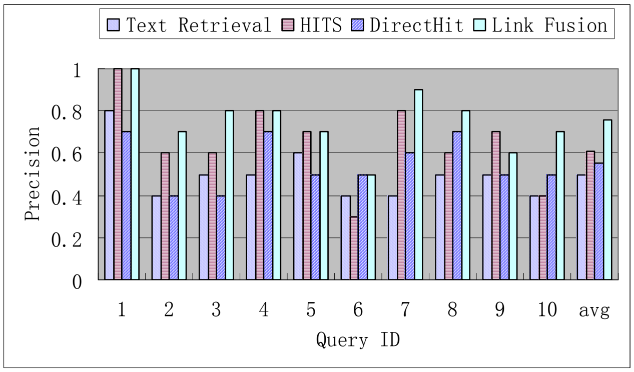

Figure 4. The precision comparison of 4 algorithms

We can see from Figure 4 that our proposed Link Fusion algorithm outperforms the basic HITS algorithm and DirectHit algorithm by 24.6% and 38.2% respectively.

4.4 Case Studies

We give a more detailed analysis of the results by looking at the top URLs returned by three algorithms for several queries. First we show the results of query "audi car" in Table 2. Shaded cells in the table indicate relevant pages. We found that the Link Fusion algorithm had returned 7 out of 9 relevant pages returned by HITS algorithm and DirectHit algorithm combined together, while only keep 2 of the 8 non relevant pages returned by HITS and DirectHit algorithm. Furthermore, Link Fusion algorithm had returned one more relevant page: http://www.s-cars.org/ that has not been found in the top 10 results from either HITS or DirectHit algorithm.

The above observations shows that the Link Fusion algorithm has the capability of keeping the correct results from different link analysis algorithms it combined, while filter out incorrect results returned from these algorithms. Researchers had reported similar findings from data fusion in information retrieval area [21]. They claimed that the combined search engine could keep the relevant results returned by different single search algorithms, while filter out those non-relevant results returned by single search algorithms. However, whether the prerequisite conditions for data fusion in information retrieval to be effective are still valid in Link Fusion problem is still left to be explored.

Table 2. Top 10 results for query "audi car"

HITS	DirectHit	Link Fusion
http://www.audiworld.com	http://www.audiusa.com/	http://www.audiusa.com/
http://www.audiusa.com/	http://www.autotrader.com/	http://www.audiworld.com
http://www.audicanada.ca/	http://www.nytimes.com/pages/automobiles/index.html	http://www.uvas.com/
http://www.vindis-cambridge.audi.co.uk/	http://pages.ebay.com/ebaymotors/browse/cars.html	http://www.s-cars.org/
http://pages.ebay.com/ebaymotors/browse/cars.html	http://www.thecarconnection.com/	http://communities.msn.co.uk/AudiSCarsUK/pictures
http://www.quattroclubusa.org	http://www.gearhead-cafe.com/mags.html	http://www.autotrader.com/
http://www.karquattro.com/	http://www.uvas.com/	http://www.quattroclubusa.org
http://www.porsche.	http://communities.ms	http://www.a4.org

com/	n.co.uk/AudiSCarsUK/pictures	/
http://www.vwvortex.com	http://www.autotrader.com/	http://www.vwvortex.com
http://www.nytimes.com/pages/automobiles/index.html	http://www.a4.org/	http://www.thecarconnection.com/

We also found that the binary relevance judgment of a web page we applied in this experiment cannot always fully reflect the "value" of a web page. Although the number of relevant pages returned within top 10 pages by the Link Fusion algorithm (8) is slightly better than that of the HITS algorithm (6), the relevant pages returned by the Link Fusion algorithm (e.g., http://www.a4.org, http://www.s-ars.org) are more authoritative than the relevant pages returned by HITS (e.g. http://www.vindis-cambridge.audi.co.uk). This problem is well represented by another case below.

Table 3. Top 10 results for query "search engine"

HITS	DirectHit	Link Fusion
http://www.google.com/	http://www.google.com/	http://www.google.com/
http://www.ubnmovies.com/	http://dailynews.yahoo.com/ fc/Tech/Internet_Portals_and_Search_Engines	http://www.excite.com/
http://www.arelanrecords.com/	http://www.search.com/	http://www.lycos.com/
http://www.novanw.com/	http://www.decideinteractive.com/	http://search.msn.com/
http://www.megaspider.com/	http://www.usaweeend.com/01_issues/010722/010722web.html	http://www.megaspider.com/
http://www.excite.com/	http://www.galaxy.com/	http://www.arelanrecords.com/
http://www.asiaco.com/	http://searchenginwatch.com/awards/	http://www.ubnmovies.com/
http://www.lycos.com/	http://www.bcentral.com/products/si/default.asp	http://www.novanw.com/
http://search.ietf.org/search/brokers/internet-drafts/query.html	http://ixquick.com/	http://www.ixquick.com/
http://www.searchenginewatch.com/	http://www.infospace.com/	http://www.dogpile.com/

Although most pages retrieved by the three algorithms are correct web pages for query "search engine", it is easy to see that the Link Fusion algorithm apparently gives higher ranks to more popular search engines (e.g., http://www.excite.com, http://www.lycos.com) than the other two algorithms. While in HITS and DirectHit algorithms, correct but not very popular search engine web pages (e.g., http://www.ubnmovies.com/, http://www.search.com/) are returned on top. This is because that if a correct web page is returned on top by the Link Fusion algorithm it must be favored by both the web editors (represented by hyperlinks) and the web users (represented by user links) rather than just one of them (e.g., HITS or DirectHit). Thus the Link Fusion algorithm returns more popular results on top than HITS and DirectHit algorithm and also more robust than the other two algorithms.

Below are the results of query "daily news". We can find from this example that the Link Fusion algorithm had both keep the correct results from HITS and DirectHit algorithm and rank the popular correct pages (e.g. http://www.nytimes.com) much higher than the other two algorithms.

Table 4. Top 10 results of query "daily news"

HITS	DirectHit	Link Fusion
http://www.surfinfo.com/html/visreport.html	http://www.msnbc.com/m/hor/horoscope_front.asp	http://www.nytimes.com/
http://dailythong.dhs.org/index.php3	http://daily.webshots.com/	http://sportsillustrated.cnn.com/
http://www.sportspages.com/regions/mw.htm	http://www.thedaily.com/bikini.html	http://encarta.msn.com/
http://www.gossipcentral.com/	http://www.poems.com/today.htm	http://www.thedaily.com/overlook.html
http://www.thedaily.com/overlook.html	http://www.alrai.com/	http://abcnews.go.com/
http://www.webcomics.com/daily.html	http://www.poems.com/	http://www.poems.com/
http://www.guampdn.com/classifieds/index.html	http://www.thedaily.com/overlook.html	http://www.thedaily.washington.edu/
http://www.nytimes.com/	http://cityguide.guampdn.com/fe/index.asp	http://www.gossipcentral.com/
http://www.thedaily.washington.edu/	http://www.thedaily.washington.edu/	http://www.thedaily.com/bikini.html
http://www.smartertimes.com/	http://dailythong.dhs.org/index.php3	http://abcnews.go.com/sections/entertainment/

5. Conclusions and Future Work

In this paper, we first defined two kinds of links among data objects within different data types: intra-type links, which represent the relationship of data objects within a homogeneous data type, and inter-type links, which represent the relationship of data objects between different heterogonous data types. Then, we proposed a unified link analysis framework, called "link fusion", to analyze inter- and intra-type links and to bring order to data objects in different data spaces at the same time.

Next, we evaluated the effectiveness of our proposed link fusion algorithm by applying it into a real world scenario of three data spaces: Hub-page space, Authority-page space, and User space. Experimental results on 10 real world sample queries show that the Link Fusion algorithm achieved 24.6% improvement over the HITS algorithm and 38.2% improvement over the DirectHit algorithm based on the measurement of precision at top 10 documents returned. After a few case studies, we found that the Link Fusion algorithm has the capability of keeping the correct answers returned by each of the link analysis algorithm it combined and trend to return the most popular results on top of its return list. These results support our assumption that the Link Fusion algorithm when used properly can help find the correct order of attributes of data objects within different data spaces.

Although the Link Fusion algorithm seems to be promising according to our preliminary experiments, there are still many issues that need to be explored. For example, in our experiment, we assumed the links from different data spaces are equally important when calculating the attributes of objects across different data spaces. However, this assumption is overly naïve, and it is almost never the case that the links from different data spaces are equally important. It is natural to think: Is there any way to identify the relative importance of links from different spaces automatically? We will explore this problem in our future research works.

REFERENCES

1. The Clever Searching, the Clever project of IBM Almaden Research Center, www.almaden.ibm.com/cs/k53/clever.html.
2. Berman, A. and Plemmons, R.J. Nonnegative matrices in the mathematical sciences. in Classics in Applied Mathematics, 1994.
3. Bharat, K. and Henzinger, M.R., Improved algorithms for topic distillation in a hyperlinked environment. in 21st ACM SIGIR International Conference on Research and Development in Information Retrieval, (Melbourne, Australia, 1998), 104-111.
4. Brin, S. and Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30. 107-117.
5. Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P. and Rajagopalan, S., Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. in 7th international conference on World Wide Web, (Brisbane, Australia, 1998), 65 - 74.
6. Chakrabarti, S., Dom, B.E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D. and Kleinberg, J.M. Mining the Web's Link Structure. IEEE Computer, 32 (8). 60-67.
7. Cohn, D. and Chang, H., Learning to Probabilistically Identify Authoritative Documents. in 17th International Conference on Machine Learning, (Stanford, California, 2000), 167-174.
8. Craswell, N. and Hawking, D., Overview of the TREC-2002 Web Track. in 11th Text Retrieval Conference, (Gaithersburg, Maryland, 2002).
9. Craswell, N., Hawking, D. and Robertson, S., Effective Site Finding using Link Anchor Information. in 24th annual international ACM SIGIR conference on Research and development in information retrieval, (New Orleans, Louisiana, 2001), 250-257.
10. Davison, B.D., Toward a unification of text and link analysis. in 26th annual international ACM SIGIR conference on Research and development in information retrieval, (Toronto, Canada, 2003), 367-368.
11. DirectHit. http://www.directhit.com.
12. Garfield, E. Citation analysis as a tool in journal evaluation. Science, 178. 471-479.
13. Hayes, B. Graph Theory in Practice, 2000.
14. Herlocker, J.L., Konstan, J.A., Borchers, A. and Riedl, J., An algorithmic framework for performing collaborative filtering. in 22nd annual international ACM SIGIR conference on Research and development in information retrieval, (Berkeley, California, 1999), 230-237.
15. Hubbell, C.H. An input-output approach to clique identification. Sociometry, 28. 377-399.
16. Katz, L. A new status index derived from sociometric analysis. Psychometrika, 18 (1). 39-42.
17. Kleinberg, J.M. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46 (5). 604-632.
18. Miller, J.C., Rae, G., Schaefer, F., Ward, L.A., LoFaro, T. and Farahat, A., Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records. in 24th annual international ACM SIGIR conference on Research and development in information retrieval, (New Orleans, Louisiana, 2001), 444-445.
19. Ng, A.Y., Zheng, A.X. and Jordan, M.I., Stable algorithms for link analysis. in 24th ACM SIGIR International Conference on Research and Development in Information Retrieval, (New Orleans, Louisiana, 2001), 258-266.
20. Pinski, G. and Narin, N. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Information Process and Management, 12. 297-312.
21. Vogt, C.C. and Cottrell, G.W., Predicting the performance of linearly combined IR systems. in 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, (Melbourne, Australia, 1998), 190-196.
22. Wen, J.-R., Nie, J.-Y. and Zhang, H.-J. Query Clustering Using User Logs. ACM Transactions on Information Systems (TOIS), 20 (1). 59-81.

Appendix:

Proof of convergence for the calculation of unified matrix A

In this appendix, we prove the convergence of the iterative calculation method of unified matrix A defined by (5). The proof of convergence is given, after the proofs of 3 lemmas.

Lemma A: A as defined by (5) is a non-negative, row-stochastic matrix.

Proof: Based on (4), we know that matrices $L_M^{'}$ and $L_{NM}^{'}$ are non-negative, row-stochastic matrix with $\alpha_M + \sum_{\forall N \neq M} \beta_{NM} = 1, \alpha_M > 0, \beta_{NM} > 0$. Thus, each element in matrix A is non-negative, and the sum of each row of matrix A is 1, which means A defined by (5) is still a non-negative, row-stochastic matrix. ■

Lemma B: If A that is defined by (5) is also reducible, there exists a permutation matrix P, such that $PAP^T = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$. Here, A_1 is a non-negative, row-stochastic and irreducible matrix.

Proof: Actually, if A defined by (5) is a reducible Matrix, then there exists a permutation matrix P, such that $PAP^T = \begin{bmatrix} A_1 & 0 \\ B & A_2 \end{bmatrix}$. Here, A_1 is a non-negative, row-stochastic, and irreducible matrix.

From the 2nd construction of the matrix defined by (5), we know, if $\beta_{MN}>0$, then $\beta_{NM}>0$. That means that if $L_{MN}^{'}$ is not a zero matrix then $L_{NM}^{'}$ is not a zero matrix either. Based on the 1st construction of matrix A, we know that L'_{MN} and L'_{NM} are all positive matrices. So A is a symmetric matrix in some sense. That is if A(*i*,*j*) is non-zero then A(*j*,*i*) is non-zero too.

Notice that the transformation of A, PAP^T, doesn't change the symmetry couple relation of A. It means that the transformed matrix PAP^T has the same feature: if element (*i*,*j*) is non-zero then the element (*j*,*i*) is non-zero. So PAP^T has the format of $\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$ ■

Lemma C If one matrix A is non-negative, row-stochastic matrix, and irreducible, then iterative calculation $x^{i+1} = A^T x^i$ converges to the principle eigenvector of A. (Assume x^0 is a positive and normalized vector).

Proof: A is a non-negative, row-stochastic matrix, also irreducible, thus, A is an ergodic transition matrix of a Markov chain MC.

According the ergodic theory of Markov chains, if we can prove that the MC has one and only one stationary probability vector x^S, then the iterative calculation $x^{i+1} = A^T x^i$ can converge to the stationary vector x^S for any initial vector x^0. Here, we assume the norm of x^0 is normalized to 1, and x^0 is positive.

To prove the Markov chain has only one stationary vector x^S, we get the following 2 points firstly:

1) For A a non-negative, row-stochastic matrix, $\rho(A)$, the spectral radius of A, is equal to 1.

2) For A a non-negative and irreducible matrix, $\rho(A)$ is an eigenvalue of A with multiplicity 1, and $\{x \mid x > 0, Ax = \rho(A)x\} = \{x \mid x > 0, A^T x = \rho(A)x\}$ [2]. Based on 2), there exists one and only one vector $x \geq 0$ (considering scaling) satisfying $xA = \rho(A)x$. From 1), $\rho(A)$ =1. Hence, there exists one and only one vector $x \geq 0$ (considering scaling) satisfying $xA = x$.

If we scale x to make the sum of x be 1, it's easy to know the equation $xA^k = x$ existed for any *k*=1, 2… So x is the stationary vector of Markov chain MC. Also, x is the principle eigenvector of A.

Hence, if A is a non-negative, row-stochastic matrix, and irreducible, then iterative method $x^{i+1} = A^T x^i$ converge to the principle eigenvector of A. ■

Theorem: For the matrix A defined by (5), the iterative method $w = A^T w$ converges to the principle eigenvector of A.

Proof: Firstly, A is a non-negative, row-stochastic matrix. If A is irreducible, then according to lemma C, we know the iterative method $w = A^T w$ converges to the principle eigenvector of A.

If A is reducible, let $w' = Pw$, here P is the permutation matrix fitting $PAP^T = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$. Then the iterative method turns out to be $w = \begin{bmatrix} A_1^T & 0 \\ 0 & A_2^T \end{bmatrix}$.

Without loss of generality, we assume A$_1$, A$_2$ are irreducible. We rewrite w' to be $\begin{pmatrix} w_1' \\ w_2' \end{pmatrix}$, then we get two sub-iterative methods: $w_1' = A_1^T w_1'$, and $w_2' = A_2^T w_2'$. Based on lemma C, these 2 methods all converge. Taking limitation on the original iterative method: $w = A^T w$, we know w is an eigenvector of A associated with an eigenvalue that equals to 1. Also, we know that spectral radius of A is 1, so w is the principle eigenvector of A. ■