Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes

Shannon Bradshaw

Department of Management Sciences
The University of Iowa
Iowa City, IA 52242
shannon-bradshaw@uiowa.edu

Abstract. Citation indexes are valuable tools for research, in part because they provide a means with which to measure the relative impact of articles in a collection of scientific literature. In retrieval systems for citation indexes, recent work has demonstrated the benefit of using ranking metrics based on measures of impact. While this approach is effective in identifying a few of the most important contributions to an area, many documents ranked highly in response to queries are irrelevant to the topic of interest. The problem here is that with such techniques Boolean methods are used to identify candidates for retrieval, even though such methods are poor determiners of relevance. As a solution to this problem, we present an indexing technique that pulls together measures of relevance and significance in a single retrieval metric. This approach, which we call Reference Directed Indexing (RDI) is based on a comparison of the terms authors use in reference to documents. Initial retrieval experiments with RDI indicate that it retrieves documents on par with significance-based techniques in terms of impact, and comparable to traditional vector-space approaches with regard to relevance.

1 Introduction

In order to contribute to a field of study a researcher must be aware of prior work related to her own and be able to appropriately position new work within that space. Citation indexes such as CiteSeer [13] have proven extremely useful in locating important research related to one's own. The ability to traverse a network of documents linked together by citations allows one to locate some of the most important contributions in nearly any research area. Furthermore, a citation index is able to "index" to some degree even papers for which it does not have access to the full text, simply because other articles in its database cite that paper. Finally, a citation index can easily identify the frequency with which an author or a specific paper is cited. Such measures are useful in determining the relative importance of documents. Recent work has demonstrated that retrieval metrics based on the impact of papers are useful means of providing researchers with at least some of the information they need [13, 9]. However, this work to date has largely left open the question of how relevance is to be determined

when ranking search results based primarily on some measure of impact. Instead, impact or significance has been used as a substitute for relevance. Currently, in most citation indexes subject search is based on Boolean retrieval, and any document using a set of query terms is an equally likely candidate for retrieval. Therefore, in any large citation index, many irrelevant documents may rank highly in the set of search results for a query, simply because they are frequently cited and happen to contain the query terms. In Web search engines such as Google, where similar techniques based on popularity are used, this approach is quite effective, given that the average information need can be satisfied by a single popular Web page. Users of citation indexes, however, do not have this luxury, because they often require an extensive treatment of a topic – information that can only be found by reviewing many documents. Therefore, users of citation indexes must often resort to the tedious process of shuffling through long lists of search results sorting the good from the bad or the equally difficult task of traversing many citation paths beginning at a few known relevant documents. This problem could be made substantially less severe if stronger measures of relevance were employed to provide users with a higher percentage of documents that are significant for what they have to say about the topic of interest. As a solution to this problem, we present an indexing technique that pulls together measures of relevance and impact in a single metric. This approach, which we call Reference Directed Indexing (RDI) is based on a comparison of the terms authors use in reference to documents. Initial experiments with this approach indicate that it outperforms Boolean retrieval, performing quite favorably when compared to a traditional vector-space approach [18] using TFIDF [19] as the term weighting metric. In addition, these experiments demonstrate that RDI selects papers that are not only relevant to a query, but those that are also among the most frequently cited for their contributions to the research area of interest.

2 Reference Directed Indexing

The intuition driving RDI is that when referencing another document, an author identifies one or more contributions made by the cited document. In doing so, he uses words that make good index terms because they identify what the document is about and the terms people typically use to identify the information it contains. For example, Figure 1 depicts a reference to a paper by Ronald Azuma and Gary Bishop entitled "Improving Static and Dynamic Registration in an Optical See-through HMD". This is an early paper on tracking the head of a user in virtual/augmented reality environments in order to present her with the appropriate perspective view for each frame of an immersive experience. Note that the citing author in Figure 1 describes this paper as addressing a six degrees of freedom optical tracking system in addition to listing details concerning its implementation. In this paper, we will refer this type of statement as "referential text", or simply a "reference". While this particular reference contains words that serve as excellent index terms, it would be difficult to build a system

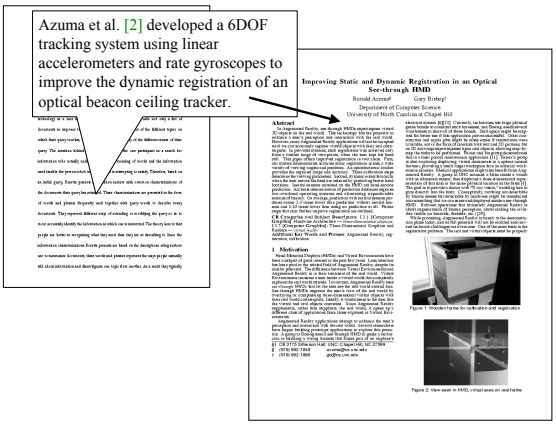

Fig. 1. The words of one referrer when citing Azuma and Bishop.

that automatically extracts just the right index terms on the basis of this text alone. Therefore, RDI leverages the fact that sufficiently useful documents are cited by multiple authors. Repeated reference to a document provides a means of comparing the words of many referrers. If several authors use the same words in reference to a document our theory is that those words make good index terms for that document.

Building on the example in Figure 1, Figure 2 depicts three additional references to the tracking paper by Azuma and Bishop.[1] In this example, each piece of text written in reference to Azuma and Bishop's paper contains many words that accurately label one or more of the key ideas in this paper, and each reference contains words also found in the words of one or more of the other authors citing this document. Note the repeated use of the underlined terms "augmented reality" and "tracking" in Figure 2. In general, the labels one citing author uses to name the contributions of a cited document are used by many other authors who cite the same document. RDI treates each citing author as a voter given the opportunity to cast one vote for any index term for a cited document. A vote of "yes" is determined by the presence of that term in the text the citing author writes in reference to the document, a "no" vote by the abscence of the term. Provided a term is not widely used in reference to many documents (i.e. articles, prepositions, and other terms that rarely identify content), the more

[1] Clockwise beginning with upper left from: S. You and U. Neumann. Fusion of vision and gyro tracking for robust augmented reality registration. In Proceedings of the IEEE Conference on Virtual Reality, pages 71-78, Yokohama, Japan, March 2001; E. S. McGarrity. Evaluation of calibration for optical see-through augmented reality systems. Master's thesis, Michigan State University, 2001; T. Auer, A. Pinz, and M. Gervautz. Tracking in a Multi-User Augmented Reality System. In Proceedings of the First IASTED International Conference on Computer Graphics and Imaging, 249-253, 1998; C.P. Lu and G. Hager. Fast and globally convergent pose estimation from video images. PAMI, 22(2), 2000.

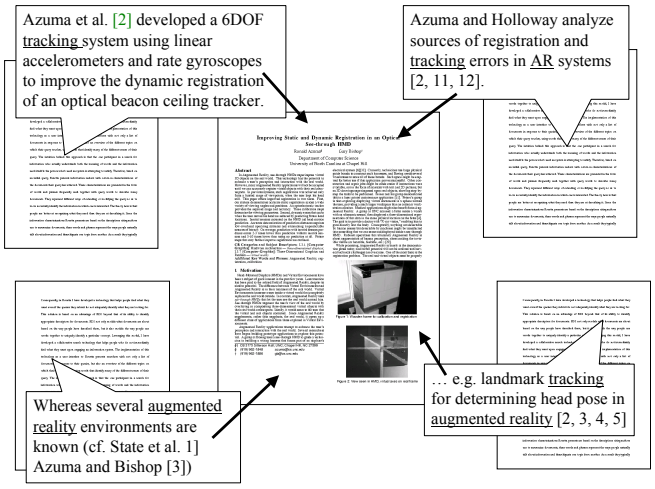

Fig. 2. Several references to Azuma and Bishop.

votes that term receives the greater its weight will be as an index term for the document. At retrieval time, then, the highest ranking documents returned in response to a query are those that have been most often referenced using the words in the query. The experimental evidence presented in Section 4 suggests that a retrieval system based on this technique provides high degree of retrieval precision, while suggesting documents that are heavily cited in the literature, and therefore, make important contributions to the topic of interest.

3 Rosetta

We implemented RDI in a search engine for scientific literature called Rosetta. For the experiments reported in this paper we indexed data provided by CiteSeer/ResearchIndex [13] with their permission. As referential text with which to index the documents in Rosetta, we used the "context of citations" provided by CiteSeer by following the "Context" link from the "Document Details" page representing each document. Each piece of referential text is approximately 100 words long, with 50 words on either side of the point of citation. Rosetta's term weighting metric is defined by:

$$w_{id} = \frac{n_{id}}{1 + logN_i}$$

where w_{id} is the importance of a word i as an index term for document d, n_{id} is the number of times word i was used in reference to d, and N_i is the number of documents for which word i is used as an index term. In response to queries, the current implementation gathers all documents indexed by the query terms and sorts them based on the number of query terms they match and

the weight of those words as index terms. The metric used to rank documents during retrieval is designed to favor documents that have been described most often using language that closely matches the query. Specifically, the score of a document is calculated as

$$s_d = n_d + \sum_{i=1}^{q} w_{id}$$

where n_d is the number of query words matching some index term for document d, q is the set of words in the query, and w_{id} is the weight of query word i as an index for document d. This metric causes documents to be sorted first by the number of query words their index terms match and then by the sum of the weights of the query words as index terms for the document. The theory here is that when citing a document, authors describe it using terms similar to those a searcher is likely to use in queries for the information the document contains. Therefore, in response to a query, the retrieval metric associates the most importance with documents that have been described using all of the terms contained in a query and then ranks search results according to the frequency with which the matching query terms have been used in reference to each document. This metric which we call "Ranked Boolean" may seem a contradiction given that we are interested overcoming weak measures of relevance provided by Boolean-based retrieval. While we acknowledge that this metric can be improved by eliminating its Boolean component, overall, as we demonstrate in Section 4 this approach seems less prone to retrieval errors than Boolean and even relevance-based retrieval techniques. We discuss this in more detail in the next section.

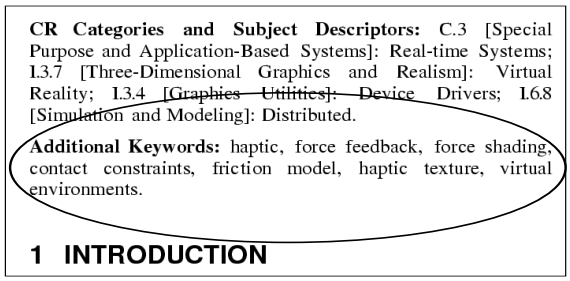

Fig. 3. We selected queries for our experiment from document keywords.

4 Retrieval Performance

We evaluated the retrieval performance of Rosetta on a small sample of 32 queries with appropriate significance tests. We simulated as realistically as possible, actual usage of Rosetta in service of research tasks. For this experiment we selected queries at random from terms used in the keywords sections of 24 documents in

our collection selected at random. For an example, see Figure 3.[2]) The queries varied in length from 1 to 3 words with 19 consisting of two words and three each of lengths 1 and 3. Note that the length of the queries used in this experiment are in keeping with the average length of queries submitted by users of both Web search engines [23] and digital libraries of scientific literature [10]. See Table 1 for the complete list of queries used in this experiment. We chose to test Rosetta

adaptive estimation	sonic feedback	groupware
supervised learning	haptic	topology changes
hardware performance counter	transient interactions	inductive transfer
reliable data distribution	information sharing	erasure codes
reinforcement learning	virtual finger	user interfaces
visual reconstruction	reliable multicast	wait free
semistructured data	wavelets	shared variables
simulation acceleration	wireless routing	wrapper induction
software architecture diagrams	code reuse	digital audio
architecture cost model	virtual environments	force shading
graphical editors	laser rangefinding	

Table 1. Queries used in experiments testing RDI's retrieval performance.

using queries collected in this way, because in order to simulate actual usage an inquiry must occur in a particular context. When people use a retrieval system to locate needed information they are motivated by a specific task or context. The context in which a user submits a query determines which documents will be useful to her in the space of all information on which her query might touch. Having selected queries from the terms authors used to describe their work, we used as context for each query the paper from which the query was drawn. We determined the relevance of each retrieved document based on whether or not it addressed the same topic identified by the query in the paper from which it was drawn. For example, one term, "reliable data distribution", was used by authors to describe research on multicast technology for distributing bulk data such as video feeds to many clients simultaneously with error detection and congestion control.[3] Using this term as a query, we marked as relevant documents that discuss multicast technology that ensures reliable distribution.

We compared the performance of RDI as implemented in Rosetta to a standard vector-space approach [18] using a TFIDF [19, 21] term weighting metric and a cosine retrieval metric [20]. We chose to compare RDI to this approach rather than a Boolean retrieval metric, because Boolean retrieval is well-known to be a poor identifier of relevant information. We felt that in the limited space

[2] From D. C. Ruspini, K. Kolarov, and O. Khatib. The haptic display of complex graphical environments. Proc. of ACM SIGGRAPH, pages 345-352, 1997.

[3] J. W. Byers, M. Luby, M. Mitzenmacher, and A. Rege. A Digital Fountain Approach to Reliable Distribution of Bulk Data. In Proceedings of SIGCOMM '98, Vancouver, Canada, August/September 1998.

available, a more convincing argument could be made by focusing on a comparison to a relevance-based retrieval technique. We compared RDI to the Cosine retrieval metric using TFIDF for term weighting, because this approach or some variant is widely used and has proven to be among the best relevance-based retrieval technologies developed by the IR community [19].

We implemented the TFIDF/Cosine system using the following term-weighting metric:

$$w_{id} = TF_{id} \cdot (log_2 N - log_2 DF_i)$$

where TF_{id} is the term frequency of term i in document d, that is the number of times term i occurs in document d. N is the total number of documents in the collection and DF_i is the document frequency of term i or the number of documents in the entire collection that contain term i [18]. As the cosine retrieval metric we used:

$$cos(d, q) = \frac{\sum_{i=1}^{T}(w_{id} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{T} w_{id}^2 \cdot \sum_{i=1}^{T} w_{iq}^2}}$$

as described in [20] where T is the number of unique terms used in a collection of documents. The magnitude of a document vector in any dimension is the weight of that term as an index for the document (w_{id}). For a term not contained in a document the weight is assumed to be zero. The weight of a term in relation to a query is w_{iq} and is in this system always equal to 1.

As data for our experiment we selected 10,000 documents from the collection maintained by CiteSeer. Each document was required to be the target of at least one citation, but otherwise the documents were selected at random. Since the RDI approach is entirely based on references to documents, this requirement guaranteed that in the experiments, both the RDI system and the TFIDF/Cosine system indexed exactly the same collection of documents.

For each of the 32 test queries we evaluated 20 search results, the top 10 from both Rosetta and the TFIDF/Cosine system. We constructed a meta-search interface that searched both systems and combined the results on a single page. The meta-search interface presented the documents retrieved in random order, with no indication of the system from which each was drawn. If a document was retrieved by both systems it was displayed only once so as not to give away its origin.

4.1 Precision at Top 10

Having evaluated the search results for each query we found that RDI compares favorably to the TFIDF/Cosine approach. In general RDI identifies documents relevant to queries with better precision, making fewer of the kind of retrieval errors common to standard vector-space techniques. The two approaches approaches exhibited largely the same pattern of retrieval, reflecting variables such as query ambiguity and coverage of each topic within the collection. However, on average Rosetta placed 1.66 more relevant documents in the top ten than the

TFIDF/Cosine system. This difference in performance is significant at the 0.1% level, with a p-value of .0007.

Rosetta performed better than the TFIDF/Cosine system for 60% of the queries and as good or better for 78% of the queries. Rosetta retrieved at least 2 more relevant documents than TFIDF/Cosine in the top ten for one-half of the queries and 3 more relevant documents or more for one-third of the queries. In contrast, the TFIDF system found 2 more relevant documents than Rosetta for only 2 queries and 3 more for only 1 query. Figure 4 depicts a side-by-side comparison over the 32 queries that comprise this experiment. A closer examination

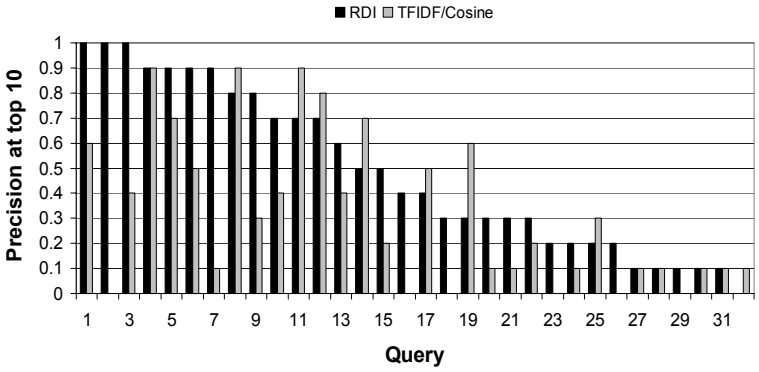

Fig. 4. Precision at top 10 for 32 queries: RDI vs. TFIDF/Cosine.

of the types of retrieval errors made by each system in this study indicates that overall RDI is less prone to many common retrieval errors than a TFIDF/Cosine approach and perhaps other content-based approaches. For example, one paper retrieved by the TFIDF/Cosine system in response to the query, "inductive transfer", had nothing to do with Machine Learning or any topics related to the query. However, this paper was retrieved with a cosine score less than 0.02 different from a very relevant document. It was retrieved because the paper contains a very lengthy banking example in which the word "transfer" is used repeatedly. Another paper, "Test Data Sets for Evaluating Data Visualization Techniques" by Bergeron et al.[4] was retrieved erroneously by the TFIDF/Cosine system in response to the query "reliable data distribution". This paper is about creating test data sets for scientific visualization applications. Because the authors discuss the appropriate distribution of values within the test data sets they create, the TFIDF/Cosine system ranked it number one in the list of search results for this query. Finally, one query used in the study was "software architecture diagrams" extracted from a software engineering paper on a formal specification for constructing software architecture diagrams. The TFIDF/Cosine system did

[4] D. Bergeron, D. A. Keim, and R. Pickett. Test Data Sets for Evaluating Data Visualization Techniques. In Perceptual Issues in Visualization, Springer-Verlag, Berlin, 1994.

not retrieve a single document directly related to this topic. Many of the papers it retrieved for this query concern software engineering topics and thus use the query words repeatedly; however, none deals directly with the topic of "software architecture diagrams". In contrast, Rosetta accurately identified four papers discussing models and tools for constructing software architecture diagrams, placing three of the four it retrieved in the top five search results. Overall, the RDI approach seems less prone to such errors. We believe this is because multiple referrers to a document seldom use many of the same words to describe it unless those words directly identify what that document is about. In contrast, the author of a document inevitably uses many words in telling the story of her research that may cause that document to be retrieved erroneously for a number of queries.

We also discovered two problems with RDI as implemented in Rosetta. We believe these two problems are to blame for the poor performance of Rosetta on queries 11, 14, and 19 depicted in Figure 4. First, although the ranking metric we have implemented performs well overall, the fact that it rewards as little as a single use of all query terms is a problem for some queries. For queries 11, 14, and 19 a few irrelevant documents were retrieved because a single reference to those documents used all the query terms even though one or more of those terms did not directly identify what the cited document was about. Therefore, we are currently experimenting with retrieval metrics that rewards documents that are frequently referenced using each of the query terms The second problem we discovered stems from the fact that many authors reuse portions of text in several papers. As a result, the same or very similar text will often appear in many pieces of referential text Rosetta uses to index a document. Because term weighting in Rosetta is based on the number of referrers that use a term, if this text contains any poor index terms, these can be a source of a false positives at retrieval time. We are currently developing a parser to detect such situations and only count the terms used in such texts a single time.

4.2 Significance of Search Results

Following our evaluation of retrieval performance, we looked next to the average number citations per year made to documents retrieved by Rosetta. In measuring the impact of documents retrieved we sought to gain some understanding of the overall significance of search results provided by the RDI approach. While for every user it is not necessarily true that a document that is frequently cited will be more useful than one that is not, it is hard to argue that a frequently cited document has not proven useful to a research community and therefore, to many people interested in the same ideas. Furthermore, similar measures of utility have been used frequently in the past in work related to our own [8, 13, 25, 11].

As a baseline for comparison we contrast the citation frequency of documents retrieved by Rosetta with those retrieved by the TFIDF/Cosine system. However, there is no reason to expect that the TFIDF/Cosine approach should prefer frequently cited documents, so we are not comparing one approach to

the other here. Rather, we do this merely to illustrate the difference between the frequency with which the documents selected by RDI are cited and the frequency with which an average document on a given topic in our collection is cited. Figure 5 depicts this comparison, graphing the median number of citations per year for the set of documents retrieved by each system in response to each query. We calculated the average number of citations per year to each

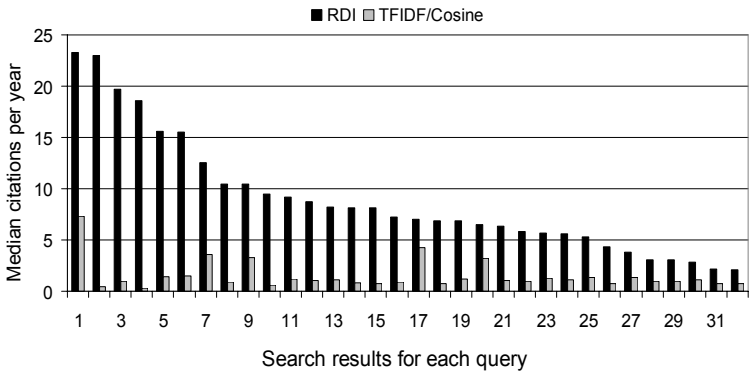

Fig. 5. Median number of cites per year to search results in our experiment.

document since its year of publication by dividing the number of years since publication by the total number of citations. The median used here then, is the median of the average number of citations per year for the set of documents retrieved for a given query. We use the median instead of mean, because it is less sensitive to a single document receiving many citations, and therefore, more reflective of the overall impact of each set of search results. We used the average number of citations per year rather than simply the total number of citations, so that the age of a document was a less significant factor in the measure of the frequency with which it is cited. As a further step in eliminating the possibility that the age of the documents retrieved by Rosetta is the cause of the difference in the number of citations, we measured the distribution of publication year for documents retrieved by both systems. We found no significant difference. The mean year of publication for documents retrieved by Rosetta is 1994, while the mean year of publication for documents retrieved by the TFIDF/Cosine system is 1995, with publication years ranging from 1984 to 2000. The frequency with which Rosetta's search results are cited far exceeds the baseline. For the average query, the median number of citations/year to search results retrieved by Rosetta was 8.9, while the baseline was only 1.5. Overall, this result combined with the study of retrieval precision indicate that an RDI approach provides search results that are both highly relevant and extremely important to the research communities to which they contribute. Though further study is required, this provides some evidence to support our claim that the RDI approach successfully combines measure of both relevance and significance in a single metric.

As a result, it is likely to provide subject search performance in citation indexes, superior to other methods previously presented.

5 Related Work

We are not the first to make use of referential text for indexing and retrieval of information; however, to our knowledge the effectiveness of the type of voting technique we propose has never been demonstrated for subject search. Furthermore, we know of work exploiting referential in relation to hypertext documents only and none directed toward scientific literature. McBryan with the World Wide Web Worm (WWWW) [14] was the first to build a search engine that incorporated anchor text. However, the WWWW provided a structured type of interface allowing users to search in anchor text as one of several choices. In addition, the WWWW provided no ranking, but simply used egrep as the underlying technology to list documents linked to using the words in the query. In other work, Spertus as a demonstration of her work in implementing structured relational database-like search of the Web built a "Parasite" tool in her SQUEAL language that successfully identified home pages using only anchor text as the basis for matches to queries [22]. Craswell et al. [4] use anchor text, which is another form of referential text, as the basis for finding homepages of companies, people, and other entities. Aside from the fact that their work is with web pages and ours with scientific literature, the primary difference between this work and ours is that while they focus on finding a specific class of documents within their domain, our approach is more generally applicable to subject searching within our domain as a whole. Other researchers have explored the idea of a reference-based approach to general-purpose indexing and retrieval to a limited extent. Bharat and Mihaila [1] use Web pages containing many links to like items as "expert pages". At retrieval time, their system identifies the expert pages most relevant to the query and retrieves the links found on those pages to which the majority of expert pages point. Of course the Google Web search engine also employs anchor text in the indexing and retrieval of web pages; however, their approach is different in that anchor serves largely just to identify candidate retrieval results, while the ranking for pages is determined primarily by PageRank [2], which is entirely subject inspecific. As a result, Google at times suffers from some of the same problems as citation indexes which rank documents based on their impact.

Others have employed referential text in classification and categorization rather in contrast to subject search. Chakrabarti et al. [3] extend HITS [12] in order to categorize Web pages into broad Yahoo-like categories, using anchor text to enhance the topic specificity of the algorithm. Furnkranz [7] uses anchor text an other elements of web pages in a Machine Learning approach to classify University web pages into one of seven categories including "Student", "Department", and "Research Project" among others.

Another body of work employing anchor text is that which uses such text to select from among several choices of links to follow. The majority of this work

deals with focused crawling of Web pages, that is, crawling with the goal of collecting information on a particular topic. Several researchers use anchor text as at least part of the basis on which candidate links are selected for the next page to crawl, acknowledging that anchor text can be a good indicator of the content of a document [24, 16, 15]. In related work, Davison [5] demonstrated that anchor text and that which surrounds the anchor text contains many terms that overlap with terms in the content of documents. In later work, published very recently [6] he used this finding as a basis for technology to guess and prefetch pages that users of Web browsers are likely to request following the page they are currently viewing.

Finally, in the general case little work has focused on techniques that merge measures of relevance and utility or significance. One notable exception is that of Richardson and Domingos [17] in which the authors present a topic-sensitive version of PageRank.

6 Conclusions and Future Work

In this paper we describe preliminary work on an indexing and retrieval technique called RDI that promises to enhance the effectiveness of subject search in citation indexes. Though the results are preliminary, they do suggest that RDI not only outperforms weak methods of relevance such as those currently employed by citations indexes, but that it actually performs favorably compared to one of the most widely used and well-established techniques for relevance ranking in retrieval systems. In future work we plan to further demonstrate the effectiveness of this technique and explore other ways in which we might leverage the precision of referential text.

References

1. Krishna Bharat and George A. Mihaila. When experts agree: Using non-affiliated experts to rank popular topics. In *Proceedings of the Tenth International World Wide Web Conference*, 2001.
2. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117, 1998.
3. S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks*, 30(1–7):65–74, 1998.
4. N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of the Twenty-Fourth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.
5. Brian D. Davison. Topical locality in the Web. In *Proceedings of the Twenty-Third International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 272–279, 2000.
6. Brian D. Davison. Predicting web actions from html content. In *Proceedings of the The Thirteenth ACM Conference on Hypertext and Hypermedia (HT'02)*, pages 159–168, College Park, MD, June 2002.

7. Johannes Furnkranz. Exploiting structural information for text classification on the WWW. In *Intelligent Data Analysis*, pages 487–498, 1999.
8. E. Garfield. *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. The ISI Press, Philadelphia, PA, 1983.
9. Steve Hitchcock, Donna Bergmark, Tim Brody, Christopher Gutteridge, Les Carr, Wendy Hall, Carl Lagoze, and Stevan Harnad. Open citation linking. *D-Lib Magazine*, 8(10), 2002.
10. Steve Jones, Sally Jo Cunningham, Rodger J. McNab, and Stefan J. Boddie. A transaction log analysis of a digital library. *International Journal on Digital Libraries*, 3(2):152–169, 2000.
11. M. M. Kessler. Technical information flow patterns. In *Proceedings of the Western Joint Computing Conference*, pages 247–257, Los Angeles, CA, 1961.
12. Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
13. Steve Lawrence, Kurt Bollacker, and C. Lee Giles. Indexing and retrieval of scientific literature. In *Proceedings of the Eighth International Conference on Information and Knowledge Management (CIKM 99)*, pages 139–146, Kansas City, MO, November 2-6 1999.
14. Oliver A. McBryan. Genvl and wwww: Tools for taming the web. In *Proceedings of the First International World Wide Web Conference*, Geneva, Switzerland, May 1994.
15. Filippo Menczer and Richard K. Belew. Adaptive retrieval agents: Internalizing local context and scaling up to the Web. *Machine Learning*, 39(2–3):203–242, 2000.
16. J. Rennie and A. K. McCallum. Using reinforcement learning to spider the Web efficiently. In *Proc. 16th International Conf. on Machine Learning*, pages 335–343. Morgan Kaufmann, San Francisco, CA, 1999.
17. Mathew Richardson and Pedro Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.
18. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
19. Gerard Salton and Christopher Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.
20. Gerard Salton and M. J. McGill. *Introduction to Modern Information Retrieval*, chapter The SMART and SIRE Experimental Retrieval Systems, pages 118–155. McGraw-Hill, New York, 1983.
21. Karen Sparck-Jones. A statistical interpretation of term specificity and its application to retrieval. *Journal of Documentation*, 28(1):11–20, March 1972.
22. Ellen Spertus. *ParaSite: Mining the Structural Information on the World-Wide Web*. PhD thesis, MIT, February 1998.
23. A. Spink, D. Wolfram, B. Jansen, and T. Saracevic. The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.
24. Michiaki Iwazume Hideaki Takeda and Toyoaki Nishida. Ontology-based information gathering and text categorization from the internet. In *Proceedings of the Ninth International Conference in Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pages 305–314, 1996.
25. J. H. Westbrook. Identifying significant research. *Science*, 132, October 1960.