NORTHWESTERN UNIVERSITY

Reference Directed Indexing: Indexing Scientific Literature in the Context of Its Use

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Shannon Bradshaw

EVANSTON, ILLINOIS

December 2002

ABSTRACT

Reference Directed Indexing: Indexing Scientific Literature in the Context of Its Use

Shannon Bradshaw

A search engine is only as good as the degree to which it provides people with useful information. Researchers in Information Retrieval (IR) have worked toward this goal by developing measures of query relevance. These techniques determine relevance on the basis of statistical measures of the frequency with which query terms are used in documents. Unfortunately, these techniques while good at measuring relevance, often poorly identify information that is actually useful. To be useful a document must be more than simply relevant to a query, it must have something of interest to say about the topic in question. In recent years, other researchers have developed techniques that determine utility on the basis of some measure of the popularity of a document. The Google Internet search engine is an example of this approach. These systems, because they are more concerned with popularity than relevance, regularly identify a few useful documents, but many that are irrelevant. Traditional IR approaches then, accurately determine relevance but not utility, while popularity approaches accurately determine utility but not

relevance. In this dissertation, I present an approach that builds on both techniques to combine measures of relevance and utility in a single metric. This technique, called Reference Directed Indexing (RDI) overcomes many of the problems with traditional IR techniques and popularity approaches. I have implemented RDI in a retrieval system for scientific literature called Rosetta. Rosetta compares multiple references to documents to determine what documents are about and the degree to which they are useful. In response to queries it provides information seekers with the documents to which the greatest number of authors have referred using the words in their query. Relying on what referrers have to say about documents has proven to be a highly effective means of determining what documents are about, and which documents on a topic are most useful. In addition to a retrieval system, I have also developed a fully automated Collaborative Query Interface (CQI) based on RDI. The CQI helps users explore an information space and resolve query ambiguity by suggesting related topics and ways of augmenting their queries.

Acknowledgements

Through the dissertation process and the graduate schooling that preceded it, I was fortunate to have help and encouragement from many people. I am most indebted to my wife, Anna, for her unwaivering faith in me and kindnesses too numerous to mention. I also wish to thank Kris Hammond for supervising and funding this work and most of all for his inspired vision of what information technology should be. Finally, many thanks and much love to the Infolab guys. To Jay Budzik for his collaboration, insight, and wisdom. To David Franklin for his friendship, collaboration, encouragement, and advice. To Josh Flachsbart, whose altruism almost makes me believe socialism can work (if the world was full of people like him). And to Robb Thomas, Sanjay Sood, Andy Crossen, and Ayman Shamma for many good times I will always remember.

Dedication

For my Dad who always shook his head and told me he didn't know how he was going to pay for my schooling. I am sure he would have been amazed to learn that in Computer Science they pay you to go to school.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

To be useful to an information seeker, a body of information must be not only relevant to his inquiry, it must also be useful. As an illustration of the distinction between relevance and utility consider an employee of a marketing firm trying to generate a certain set of statistics from a spreadsheet in Microsoft Excel. If the employee formulates his question as a query and submits it to a search engine, a newsgroup posting asking the same question would be relevant. However, a page containing an answer to his question would be useful. For most of its existence the holy grail of work in the field of Information Retrieval (IR) has been search technology that ranked all documents relevant to queries higher than all non-relevant documents. After many years of research in this area most systems in wide-spread use are based to one degree or another on a vector space model [**47**] in which queries and documents are represented as vectors of terms in an n-dimensional space. In response to a query, these systems rank highest those documents whose vectors are most similar to the query vector. Systems based on these standard IR techniques while often presenting search results that are quite relevant to queries, continually retrieve information that is not very useful because many of the relevant documents they retrieve are not as important to a topic of

interest as others judged less relevant solely on the basis of a statistical analysis of word use.

More recently, with the introduction of HITS [**34**], PageRank [**9**] and subsequently Google, many researchers have turned their attention toward information technology that substitutes measures of popularity for measures of relevance. While such techniques often provide information seekers with one or two very valuable pieces of information, much of the result set for any query represent documents that are quite useful, but for queries other than the one for which they were retrieved. This is no less true of some retrieval work in the domain of scientific literature, where for decades researchers in bibliometrics [**32, 58**] and citation analysis [**26, 36**] have measured the authority or significance of bodies of work using measures involving the frequency with which that work is cited. One of the most visible bodies of work in this area currently is that of ResearchIndex, formally known as CiteSeer [**36**]. ResearchIndex automates citation indexing as originally conceived by Eugene Garfield [**27**]. It constructs explicit and navigable networks of research articles from the implicit networks defined by citations between them. While such techniquesIn ResearchIndex a user can traverse links to both articles cited by that document and articles that cite that document and continue traversal throughout the network. Though the primary contribution of ResearchIndex is the construction of traversable networks of research papers, Lawrence et al. have made some effort to implement document ranking based on significance. ResearchIndex gives users the option of ranking documents based on the number of citations they

receive in a year. Since citation frequency is a good measure of the importance of a research paper, by ranking documents in this way information seekers can easily find documents that make significant contributions to their respective research areas. While this method of ranking search results does promote papers that have made significant contributions to a body of knowledge to the top of the results list, many of the search results are off-point, because the system retrieves all documents containing the query terms even once. As you might imagine many of these have little or nothing to do with the topic of interest. For example in a recent search for a corpus of documents from the MEDLINE database of journal articles published by the U. S. National Library of Medicine [**42**], I submitted the query "medline and corpus" to ResearchIndex and received the results partially pictured in Figure 1.1 in response. The first document in the list of search results was somewhat useful. However, as was the case with many of the documents retrieved in response to this query, the next two merely refer to MEDLINE in passing; having nothing whatever to do with a MEDLINE corpus. Although many of these articles use the term "MEDLINE" only once, they were ranked highly because many other articles have cited them. Even though the reason why they were cited has nothing to do with what they have to say about a MEDLINE corpus. As another example, in another recent query, I searched ResearchIndex for information on web spiders that pick their path through the Internet intelligently rather than using simple breadth-first search. I searched ResearchIndex using the query "intelligent and web and spider". Again many of the top search results were completely irrelevant,

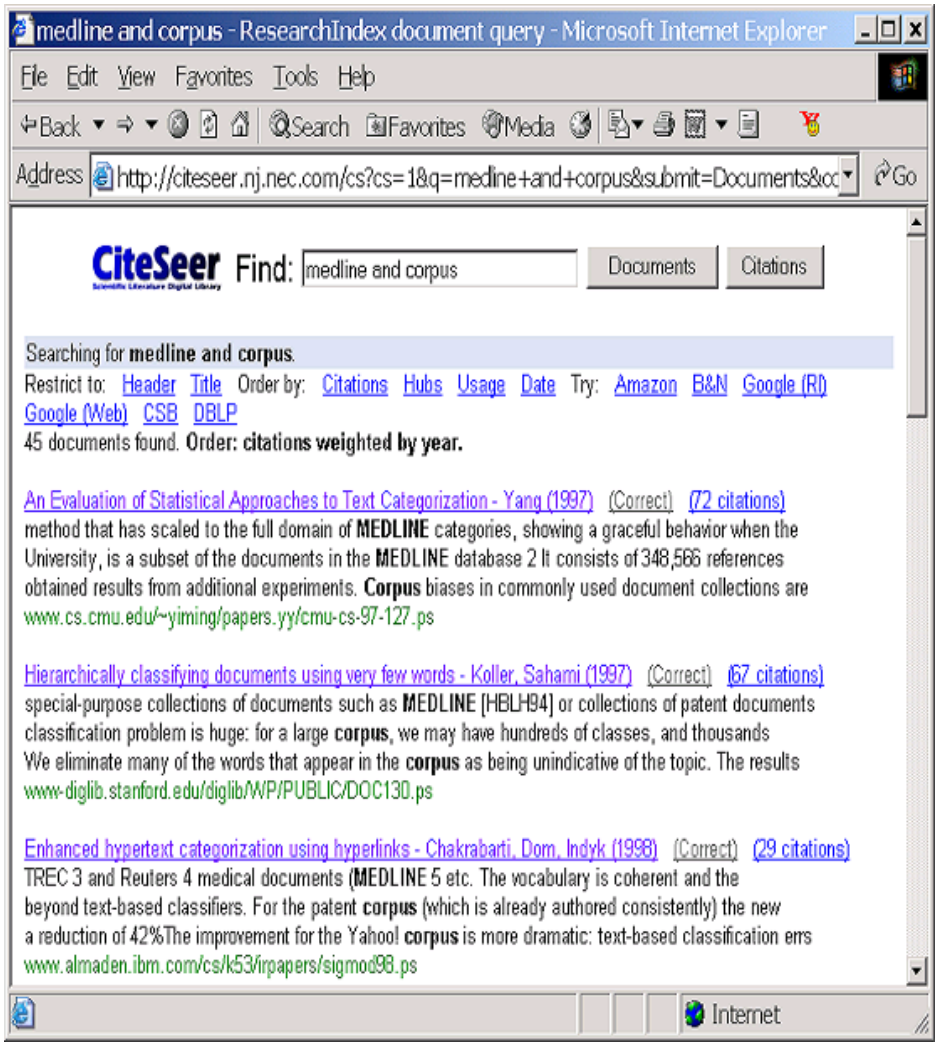


Figure 1.1. Many of the documents retrieved by a query to ResearchIndex, "medline and corpus", merely mention MEDLINE once in passing, not addressing any conceivable sense of the query.

but cited frequently. Of particular interest in this query were the documents pictured in Figure 1.2. Both of these papers, though entirely irrelevant were retrieved in the top ten search results even though the only usage of one or both of the

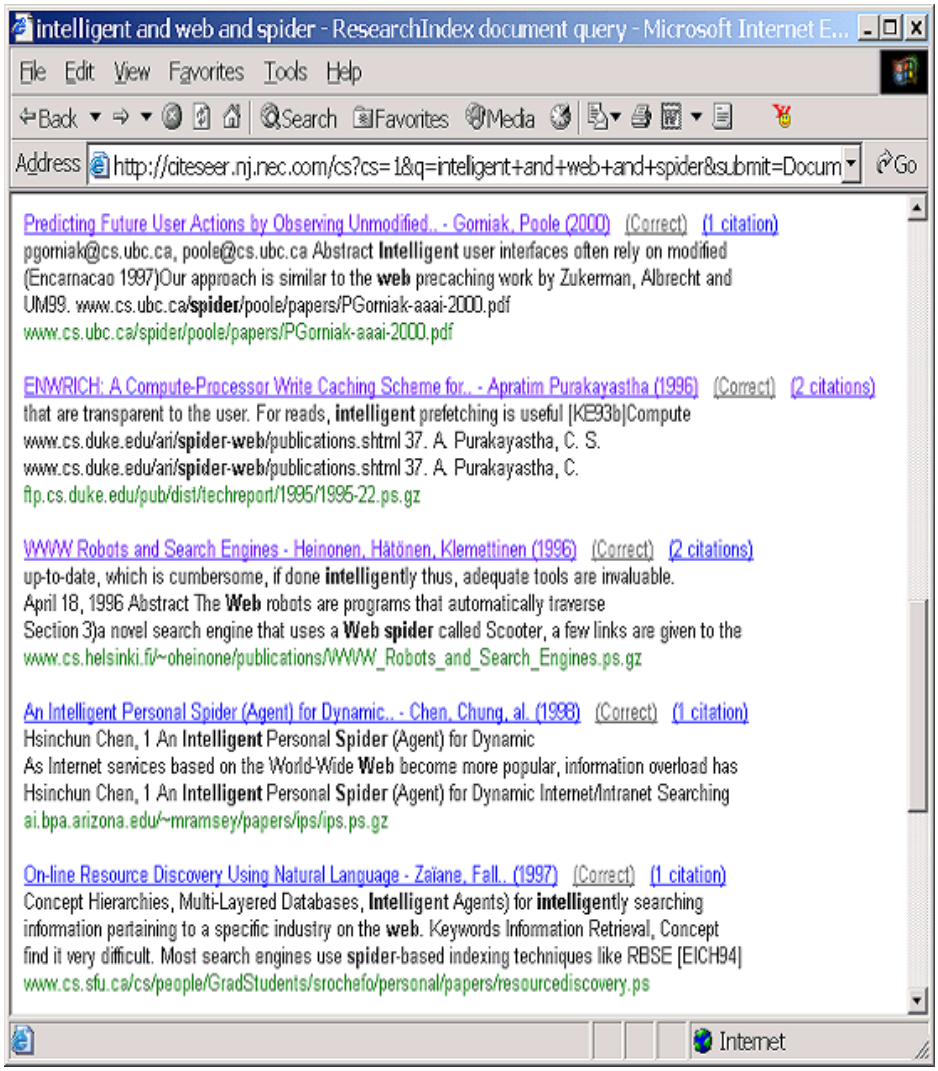


Figure 1.2. Two examples of a common problem with Boolean retrieval. Although the query, "intelligent and web and spider" and these two documents overlap largely only through words used in URLs associated with them they are retrieved within the top 10 documents for this query.

words “web” and “spider” in relation to these documents is in URLs with which they are associated. I must stress that I do not include this criticism as an attack of ResearchIndex. Retrieval is not the focus of this work and the developers make no claims about the effectiveness of the search functionality the system provides; instead, the focus is on autonomous citation indexing, an effort with which they have been very successful. I use ResearchIndex as an example merely because it is the most visible of the type of work done to include some measure of significance in the retrieval of scientific literature.

The problems pertaining to relevance exhibited by link-analysis approaches are not unique to information retrieval (IR) systems for scientific literature. They also arise in the techniques designed for the Web such as PageRank. In general the techniques developed to date that are designed to recognize the importance of information sacrifice relevance in the pursuit of popularity. As a result, indexing and retrieval technologies fall largely into two primary categories: those that rank documents on the basis of their relevance [**47, 50, 21, 29**] and those that rank documents on the basis of their popularity [**9, 34**]. Building on the relative successes of both approaches, in this dissertation I present a search technology that integrates measures of relevance and popularity as a means of determining the utility of documents retrieved in response to queries. This technique uses not simply the links formed by citations, but also the text surrounding those citations. This type of text has been called the context of a citation in other work [**36**], but I will identify such text using the term “referential text” or simply “reference”.

I will refer to the indexing and retrieval techniques presented here collectively as Reference Directed Indexing (RDI). The intuition driving RDI is that when an author cites a document, at the point in the body of the paper where the reference is made, he indicates which ideas in that document are relevant to his own research. Stated another way, he identifies one or more contributions made by that document. In doing so, he uses words that make good index terms because they identify what the document is about and the words people use to identify the information it contains. For example, Figure 1.3 depicts a reference to a paper by Azuma and Bishop entitled "Improving Static and Dynamic Registration in an Optical See-through HMD" [1]. This is an early paper on tracking the head of a user in virtual/augmented reality environment in order to present him with the appropriate perspective view for each frame of an immersive experience. Note that the citing authors describe this paper as addressing a six degrees of freedom optical tracking system in addition to listing details concerning the implementation of this system. While this particular piece of referential text is densely packed with words that serve as excellent index terms, there is no general-purpose solution that would permit an indexing system to recognize that fact on the basis of this piece of text alone. As a solution to this problem, and one that has proven quite effective [**7**], RDI leverages the fact that sufficiently useful documents are cited many times. Repeated reference to a document provides a means of comparing and contrasting

[1]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.

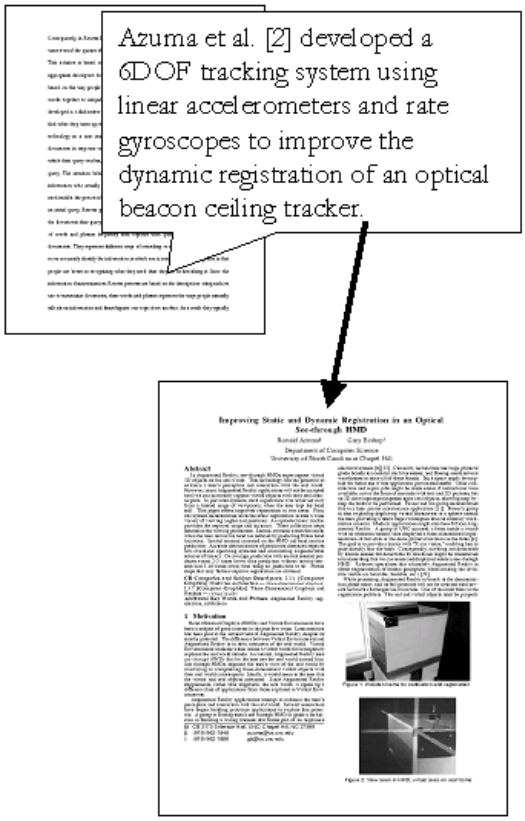


Figure 1.3. The words of one referrer when citing Azuma, R. and Bishop, G. Improving Static and Dynamic Registration in an Optical See-through HMD. In Proceedings of SIGGRAPH '94, pp. 197-204, 1994

the words of multiple referrers. If many different authors use the same words in reference to a document then it is usually the case that those words make excellent index terms for that document. Building on the example in Figure 1.3, Figure 1.4 depicts three additional references to the tracking paper by Azuma and Bishop.[2]

[2]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

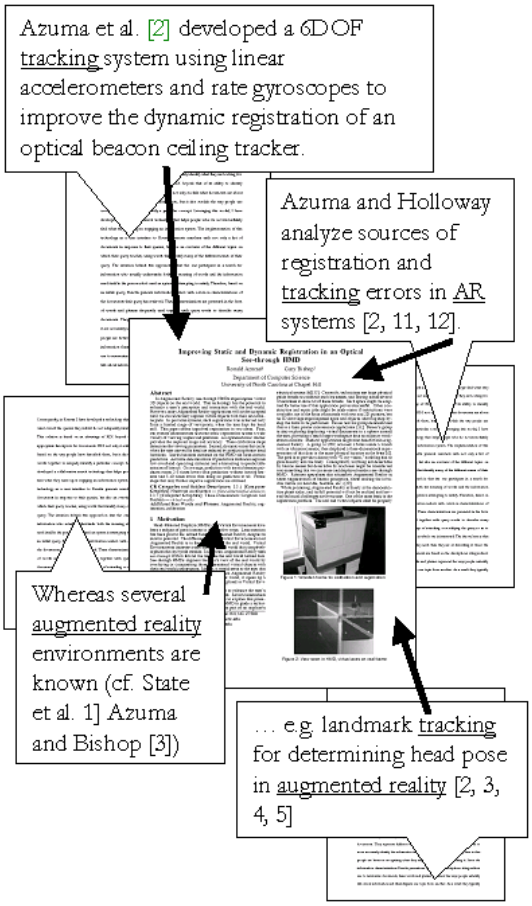


Figure 1.4. The words of several referrers when citing Azuma and Bishop. Note the repeated use of the terms "augmented reality" and "tracking".

Each piece of text in this example written in reference to Azuma and Bishop's paper contains many words that accurately label one or more of the key ideas in

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.

this paper, and each snippet uses words also found in the words of one or more of the other authors citing this document. In this example, the terms "augmented reality" and "tracking" have been used by at least two authors in describing this document. A brief look at additional references to this paper highlights other important ideas discussed and reinforces the two I have already identified. For example, those depicted in Figure 1.5, emphasize the terms "head mounted display" and "optical" and reinforce "augmented reality".[3] In this research area, the term "optical" distinguishes the type of head tracking Azuma and Bishop describe in this paper from others such as those employing magnetic sensors as a means of determining the location of a users head in relation to a fixed reference point. Finally, the referrers whose words are represented in Figure 1.6 point out yet another distinguishing feature of Azuma and Bishop's approach.[4] That being that their technique is predictive in that it attempts to infer where the users head will be in the next few frames so that the appropriate views can be pre-computed to enable a more smoothly displayed immersive experience. These references in addition to the others leave us with "augmented reality", "head mounted display", "tracking", "optical", and "predictive" as the terms that appear to most accurately describe

[3]Clockwise beginning with upper left from: M. Bajura and U. Neumann. Dynamic Registration Correction in Video-Based Augmented Reality Systems. IEEE Computer Graphics and Applications, 15(5):52–60. 1995; R. Whitaker, C. Crampton, D. Breen, M. Tuceryan, and E. Rose. Object Calibration for Augmented Reality. Proc. EUROGRAPHICS'95, pp. 15-27, 1995; D. LaRose. A Fast Affordable System for Augmented Reality. Master's thesis, Carnegie-Mellon University, 1998.

[4]From M. Bajura. Merging Real and Virtual Environments with Video See-Through Head-Mounted Displays. Doctoral Dissertation, The University of North Carolina, 1997 (top) and M. Bajura and U. Neumann. Dynamic Registration Correction in Video-Based Augmented Reality Systems. IEEE Computer Graphics and Applications, 15(5):52–60. 1995. (bottom)

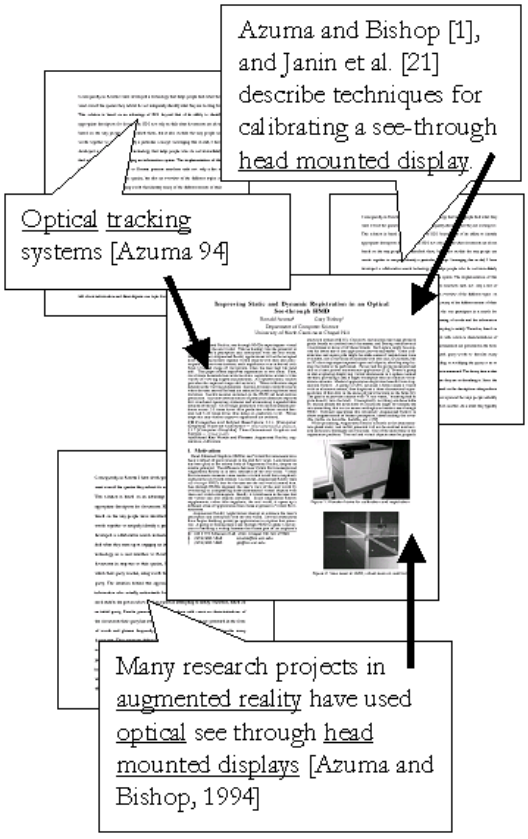


Figure 1.5. More references to Azuma and Bishop. Note the repeated use of the terms "optical", "head mounted display", and the use, once again, of the term "tracking".

what this document is about. Reading this paper, one quickly realizes that these terms do in fact identify most of the primary ideas Azuma and Bishop address. (References not included in these examples hit the rest.) This is type of result is not limited to Azuma and Bishop's paper; rather most referrers write about the documents they cite in such a way that when their words are compared with even just a few others, many of the best index terms for a document may be easily

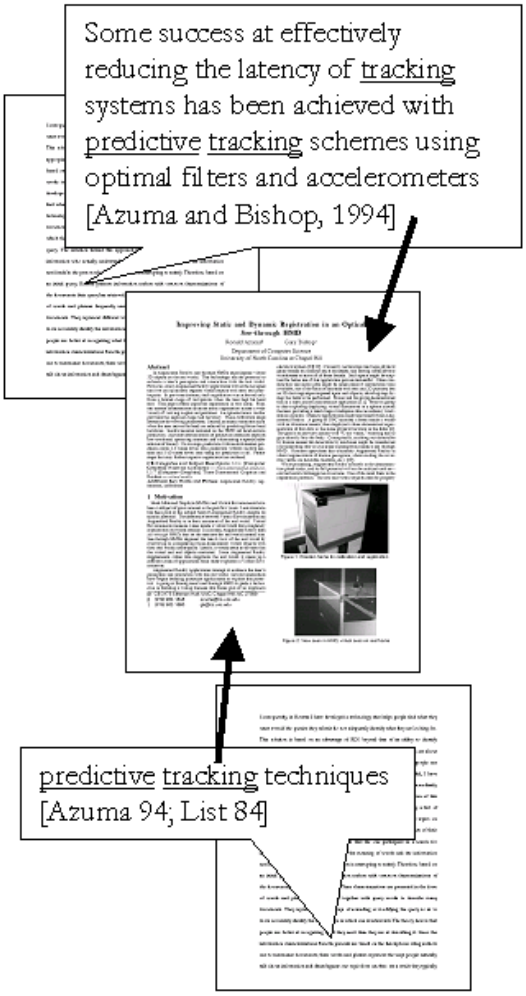


Figure 1.6. Still more references to Azuma and Bishop. Note the repeated use of the terms "predictive" and "tracking" or "predictive tracking".

identified and appropriately weighted using only shallow techniques. Furthermore, since these words have been selected from examples of the way people name this information, they are likely to match the words other people will use in queries for

the same information. Of course, there is a small problem here in that many people describing the same idea tend to use many different words to describe it [**24**], but because RDI selects index terms from the words of many different people, it not only selects the index terms that best describe the ideas addressed in a document, but also several terms for many of these ideas. I will address this issue in more detail, providing experimental evidence to support this claim in Chapter 4.

Leveraging this property of repeated reference to documents, I have developed an information retrieval system called Rosetta that through a statistical analysis of the words used in many references to a document identifies the words people most frequently use to describe that document. As a result, Rosetta precisely identifies the index words that most accurately identify what a document is about. In addition, Rosetta's retrieval ranking metrics promote documents that have been most often described using the words in a query. Stated another way, Rosetta locates the documents that people most often refer to when discussing the topic identified in a query. As a result, Rosetta provides people with documents that are both highly relevant and very important in what they have to say about the topic of interest.

Other researchers have employed referential text to one degree or another in search systems [**41, 18, 9, 14, 16**]. The most prominent example is the Google search engine as described in [**9**] uses anchor text as one source of index terms for documents. However, the use of referential text in this work differs significantly from the way it is used in this system. While Google simply uses anchor text as

another source of index terms among many, returning search results ranked primarily on the basis of their popularity, Rosetta compares and contrasts multiple references to a document to determine not only which words most accurately describe that document, but also which documents people most often refer to when talking about a specific topic. In doing so Rosetta overcomes many of the shortcomings of both traditional IR techniques and link-analysis approaches to information retrieval in that it embodies a unified technique that provides information seekers with search results that are both highly relevant and significant in what they have to say about the topic of interest. In fact, my research indicates that a reference-based approach to indexing not only provides more significant search results, but also a higher number of relevant documents on average. Indeed, a key theory on which RDI is based is that the words used within a document are a poor representation of what that document is about and queries, which are descriptions of information needed, are more accurately matched to references, which are descriptions of information provided. People who use information retrieval systems in a world where they have become ubiquitous are for the most part unfamiliar with any but the most basic principle on which these systems are based. To search for information most people simply type two or three query words, ignoring any advanced search options, even those as simple as placing quotation marks around sequences of words intended as phrases [**55**]. Searchers typically submit no more than a few queries of this type for each information need and then either find what they want or give up looking. Additional research suggests that this behavior is

not isolated to the Web but occurs even among the most technologically savvy of all users of information technology. Jones et al. [**30**] in their study of search behavior of users of the New Zealand Digital Library found that most searchers are unwilling to do more than type short simple queries even when provided with advanced search capabilities, even though the users of these libraries should be some of the most savvy of all users of information technology. With this type of search behavior, it is often difficult to match queries to documents on the basis of their content. For example, returning to Azuma and Bishop's paper on head tracking technology for augmented reality systems, many documents on both augmented and virtual reality applications contain a discussion of tracking the user through the virtual space. In the process, they use many of the words a searcher might use in queries for AR/VR papers that specifically address tracking technologies. As a result such information is difficult to find. In fact, this problem is common to many of the topics about which people search for information. Many of the words people choose to distinguish the information for which they are looking are common to the space as a whole and are therefore used throughout many documents related to the subject of their inquiry. As a result information retrieval technology based on the words used within documents often performs poorly. In contrast, references to documents are often concise descriptions of the information those documents contain, descriptions that are easily and effectively matched to queries for the same information. Such descriptions are less likely to contain misleading index terms

than are the documents themselves. In a small way this is due to the great difference in the number of words used in a sentence or two composing a reference when compared to an entire document. But a more important factor is the difference in the intent of the two types of text. A document is intended to convey one or more complex ideas using as many words as are necessary to promote understanding. As a result, documents often contain a great deal of background material, quotations from other sources, examples, and other components that can easily mislead automatic indexing and retrieval systems. The sentences in which an author refers to another document are typically constructed in such a way that they identify the idea or ideas that are germane to the topic on which she is writing. After all it is from these brief descriptions that we as researchers often locate work we should know about. The primary advantage of the RDI solution, however, is the fact that repeated reference to a document permits the use of simple statistical techniques to identify the words that most accurately identify what that document is about. A demonstration of which is a contribution of this dissertation in its own right.

Taking this work farther I go beyond simple search and begin to look at next generation information technology that helps to alleviate the difficulty with which many people construct queries that adequately describe their information needs. While the indexing techniques implemented in Rosetta improve the success with which users of digital libraries will be able to find what they need using simple natural descriptions of the information they need, there remain many types of information people will continue to have difficulty describing and therefore, finding.

The reason for this is that there is a fundamental problem in requiring people to generate queries for the information they need. The problem is, quite simply, that creating a list of words that unambiguously identify the information one needs is very difficult, and the problem becomes more severe as the information needed becomes more complex, specific, or obscure. Consequently, in Rosetta I have developed a technology that helps people find what they want even if the queries they submit do not adequately identify what they are looking for. This solution is based on an advantage of RDI beyond that of its ability to identify appropriate descriptors for documents. RDI not only models what documents are about based on the way people have described them, but it also models the way people use words together to uniquely identify a particular concept. Leveraging this model, I have developed a collaborative search technology that helps people who do not immediately find what they want upon engaging an information system. The implementation of this technology as a user interface to Rosetta presents searchers with not only a list of documents in response to their queries, but also an overview of the different topics on which their query touches, using words that identify many of the different senses of their query. The intuition behind this approach is that the one participant in a search for information who actually understands both the meaning of words and the information need itself is the person who's need an system is attempting to satisfy. Therefore, based on an initial query, Rosetta presents information seekers with common characterizations of the documents their query has retrieved. These characterizations are presented in the form of words and

phrases frequently used together with query words to describe many documents. They represent different ways of extending or modifying the query so as to more accurately identify the information in which one is interested. The theory here is that people are better at recognizing what they need than they are at describing it. Since the information characterizations Rosetta presents are based on the descriptions citing authors use to summarize documents, these words and phrases represent the ways people naturally talk about information and disambiguate one topic from another. As a result they typically draw meaningful distinctions between the different senses in which the words of a query have meaning. In recent related work, researchers have begun to refer to this type of technology as an interactive retrieval interface [**3, 38**]. Since interactive retrieval interfaces often include a relevance feedback component [**46**] and mine does not, I will forgo the use of that term and instead refer to this technology as a Collaborative Query Interface (CQI). The CQI implemented in Rosetta provides a good overview of the various meanings of a query and gives people the ability to navigate to their goal in an efficient and effective manner. This technology transforms the difficult task of generating an effective query into a comparatively simply recognition task.

In summary, the work I contribute in this dissertation centers on an exploration of the value of referential text as a means of indexing and retrieving information. Since so much of the information with which we now interact exists in an on-line form, making the creation of direct and clickable references to related information trivial, this type of text is created by everyone from Bill Gates to my Mother on a

daily basis in the natural course of writing. Therefore, the findings presented here may very well have far-reaching application to many forms of information both now and in the future.

The organization of this dissertation is as follows. In the next chapter, I describe Rosetta, detailing the implementation of RDI in that system. In Chapter 3, I present a performance evaluation of Rosetta, a study that indicates that RDI overcomes many of the problems with both content-based and link-analysis techniques for information retrieval in digital libraries, providing search results that are both on-point and among the most distinctive with regard to their treatment of the query topic. In addition, this study indicates that RDI may actually retrieve documents relevant to a query with greater precision than traditional content-based indexing and retrieval techniques. In Chapter 4, I demonstrate that a reference based indexing solution more accurately identifies what documents are about insomuch that it chooses better index terms for documents than does a content based solution for the same collection. In Chapter 5, I describe a CQI that leverages the fact that reference provides not only precise descriptions of the information documents contain, but also an accurate model of the language people use to distinguish one topic from another. Chapter 6 contains a discussion of lessons learned through this research. In Chapter 7 I discuss related work; and in Chapter 8 my plans for future work. Overall, the reference-based information access techniques

I present in this dissertation are simple, general purpose techniques which experimental evidence suggests should be considered in the development of many types of information retrieval systems.

CHAPTER 2

Rosetta

In this chapter, I describe the implementation of Rosetta, an information system I built to develop and demonstrate reference-based indexing. Rosetta is a Web-based search engine for scientific literature. It uses the words of citing authors to index each document in its collection. In developing Rosetta it was not my goal to develop complex indexing and retrieval algorithms. Rather with this system I intend to demonstrate the value of the RDI approach through the effectiveness of the simple techniques Rosetta employs. The name Rosetta is borrowed from the Rosetta Stone (See Figure 2.1). This artifact, found in Egypt in the late 18th century, contains a proclamation from an ancient ruler of Egypt. The stone in interesting because it contains text from three languages: Egyptian Hieroglyphic, Egyptian Demotic, and Koine Greek. At the time of its discovery Egyptologists did not understand Hieroglyphic; however, they could read both Demotic and Greek. The three pieces of text, the Demotic, Greek, and Hieroglyphic are linked by virtue of their appearing on the same piece of stone. In an attempt to decipher the Hieroglyphic those who studied the stone compared the text of the Demotic and Greek. Finding a great deal of overlap between the two messages, archaeologists were able to use this finding to decipher the message of the third piece of text, the

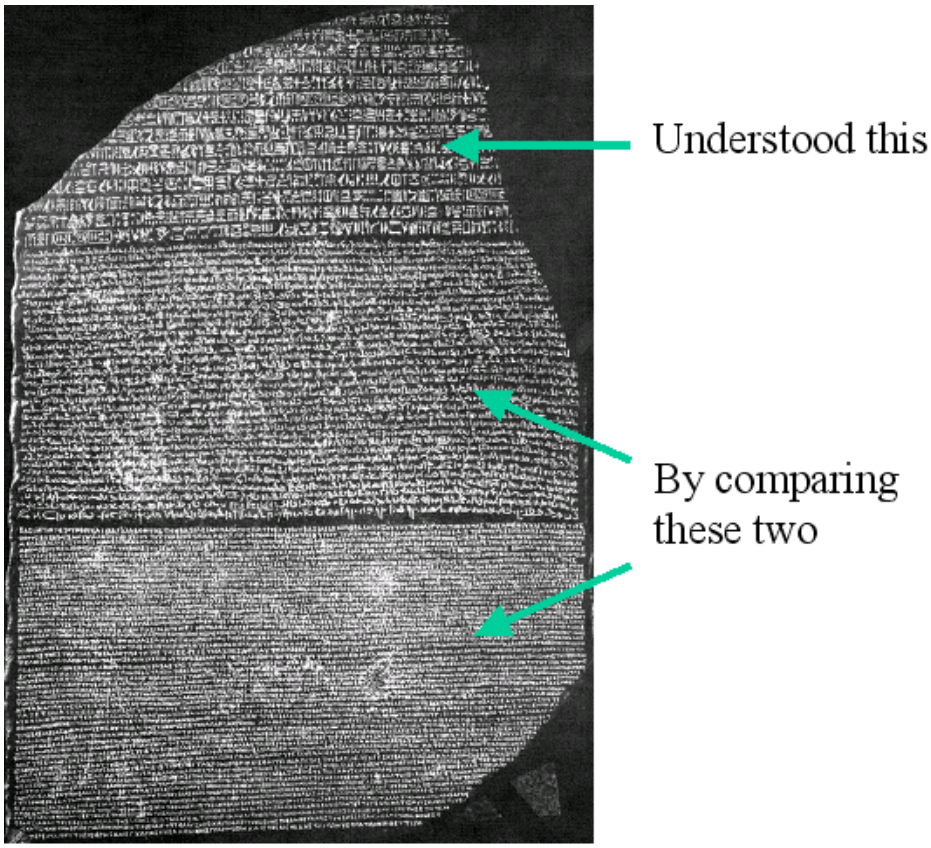


Figure 2.1. The Rosetta Stone from which the Rosetta search engine takes its name. The Rosetta stone was the artifact that lead researchers to an understanding of hieroglyphics. The catalyst for this understanding was a process of comparing the Demotic and Greek language texts, finding similarities, and using this similarity to understand the content of the hieroglyphic script.

hieroglyphic. While the process was actually one of deciphering the meaning of individual hieroglyphs, in the abstract the Rosetta Stone provides the appropriate image of the RDI approach to indexing in that two pieces of text, linked to a third are compared and contrasted to decipher the information content of the third.

2.1. Rosetta's Index

The test collection indexed by Rosetta and used in experiments presented later represents a portion of the collection maintained by ResearchIndex [**36**]. Steve Lawrence of NEC Research was kind enough to make this data available to me. Rosetta currently indexes documents solely based on the descriptions of citing authors. Each piece of referential text used to index documents in this collection is composed of a window of approximately 100 words, 50 on either side, surrounding the point at which a citation to a document occurs. These windows capture roughly the sentence containing a citation as well as the sentences before and after. The following is an example of the type of text Rosetta uses to index documents:

> ...information) provides a suite of DTDs for encoding basic document structure and linguistic annotation, and specifies a corresponding data architecture for linguistic corpora. The eXtensible Markup Language (XML) is the emerging standard for data representation and exchange on the World Wide Web (Bray, Paoli, Sperberg-McQueen, 1998). Although at its most basic level XML is a document markup language directly derived from SGML (i.e. allowing tagged text (elements) element nesting, and element references) various features and extensions of XML make it a far

> more powerful tool for data representation and access. For example,...[1]

Using this text, Patrice cites Bray, Paoli, and Sperberg-McQueen's paper, "Extensible Markup Language (XML) Version 1.0". In Rosetta, only documents cited at least once are accessible to information seekers. In one respect this is an advantage to the RDI approach, given that in order for information seekers to see a document in a set of search results, it must go through a certain amount of vetting by the community in which it is published. Furthermore, as a quick perusal of a citation index such as ResearchIndex indicates, little work of interest remains uncited for long; many receive several citations during the first year following publication. However, in order to provide a more complete solution Rosetta should make accessible other documents that have not been cited. I have begun to consider ways in which Rosetta might be augmented to provide access to documents that have not been cited, in particular to capture recent publications that haven't yet been digested by the community. Not wishing to degrade either the relevance or significance of search results, one approach might be to implement a citation index so that from any document listed in a set of search results, information seekers might explore the list of documents citing that document. The intuition here is that one can often find useful information in looking at the documents that cite some of the more important work in the areas in which he is interested. Another approach,

[1]From N. Ide, P. Bonhomme, and L. Romary. 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora. In Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association, 825-30.

since I am primarily concerned with missing recent publications, would be to index recent, uncited publications by content and present them separately as "More recent publications in which you might be interested." Finally, it may well be the case that an integrated approach in which combining RDI and more traditional index may improve recall of uncited but important documents, while having no negative effect on the relevance and significance of search results. However, in this dissertation I wish to contribute a thorough exploration of the benefits of indexing documents on the basis of referential text, since no such exploration has yet been published in either work on scientific literature or the Web. Therefore, other approaches such as those I have just outlined, while important considerations, are beyond the scope of this dissertation.

2.2. Referential Text

Rosetta uses the words found in texts such as this to index the cited document. For indexing data I used "contexts of citations" collected by ResearchIndex. Though ResearchIndex indexes documents by content, it extracts the text surrounding a citation so that users may view the ways in which authors have cited documents. This is often helpful in determining whether or not a document will be useful. Most of the language captured by the text windows Rosetta uses does describe a single cited document; however, in many references some of the text identifies other ideas. Typically this problem arises when multiple documents are cited within a single reference. By multiple citations I do not mean references such

as the following in which several documents are cited for the same reason and thus using the same description.

> ...and can perform poorly when sending compressed data formats. Nevertheless, a more extensive comparison would be interesting, particularly relating to energy efficiency. Several studies have been performed in which the browser itself is actually split across the client machine and base station [4, 9, 18]. The potential for optimization in this case is much greater since the base station can perform content specific lossy compression such as scaling and dithering of images before sending them to the mobile client. The base station side of the browser can also handle the fetching of embedded ...[2]

Rather, I find this problem in references that contain citations to multiple documents where each is described using different text. For example,

> ...standardization effort has been called for to develop very low bit-rate audio visual coding. This coding scheme called MPEG 4 is to be used on wireless mobile networks such as PCS networks. Recent research on mobile computing concentrated on improving the performance of reliable transport protocols [7] running on top of a network layer protocol such as mobile IP [17] The indirect

[2]From R. Krashinsky. Efficient Web Browsing for Mobile Clients using HTTP Compression. Term Project. Massachusetts Institute of Technology. December 2000.

> TCP or I TCP is proposed to provide transport layer communication between mobile and fixed hosts [3] In [4] this technique is further improved. The investigations were made mainly for the wireless local area networks ...[3]

In this reference the phrases "reliable transport protocols", "mobile IP", and "indirect TCP" among many others all identify different bodies of work. In references such as these an author typically ties together several documents using some thread of similarity. In doing so he uses some language that classifies them as a group and some that identifies them individually. A source of similar problems are windows of text used as references that are simply too large, capturing language that does not identify the information contained in any cited documents. For example in the following reference:

> ...In fact, browsers do not parse the HTML sources they access, and even in the presence of errors, they manage to display the corresponding page anyway. An alternative is represented by the adoption of procedural languages such as Perl [43] or Python [6]. Indeed, languages of this class behave well in the presence of exceptions, which can be explicitly caught in the control-flow of

[3]From B. Sarikaya. Multimedia Communication in Wireless Networks with an Evaluation of Slot Aggregation in PACS.

> the language, and managed separately on the basis of a case-by-case analysis. However, they have the usual...[4]

The first sentence identifies an idea that is entirely different from the second sentence containing the citations to two documents on the programming languages Perl and Python respectively. The third sentence mentions some properties of Perl and Python, but not properties that are not explicitly discussed in either of the cited documents. Although I have successfully implemented parsers that correctly handle many misleading references [**6**]. The implementation of Rosetta used in studies presented in this dissertation does not attempt to eliminate unwanted text beyond stop words and character strings that do not meet the requirements for index terms, because I wanted to test the most basic implementation of RDI and in so doing demonstrate that the technique is able to overcome a large volume of noise in the references it uses to index documents. I built this system as a general solution that merely accepts a piece of text and a pointer to the document it references, and indexes the document without any domain-specific processing. My purpose in this is to argue for the broad applicability of the findings presented in later chapters for collections of literature such as hypertext in which references from one document are common and are used in a fashion similar to that of citation in scientific literature. I seek to demonstrate that regardless of the type of document, repeated use of the same words in reference to that document by many

[4]From V. Crescenzi and G. Mecca. Grammars have exceptions. Information Systems, 23(8):539–569, 1998.

different authors provides the information necessary to provide a better basis for indexing than content, even if references contain a many noise words. Therefore, Rosetta indexes documents using all valid index terms found in references to them.

2.3. Term Weighting and Retrieval Ranking

As an additional effort in demonstrating the power of even the simplest RDI approach, Rosetta indexes documents on the basis of individual words, making no attempt to identify phrases. To be considered a valid index term, a word may contain only alphanumeric characters, hyphens, and periods. They must also begin and end with an alphanumeric character. Many useful document identifiers contain both letters and numbers, for example "B2B"; several others contain periods (i.e. "java.lang.String"), and of course, many contain hyphens (i.e. client-server). Rosetta indexes the documents in its collection using words extracted from text used in reference to them that meet the restrictions outlined above. In order to associate the best documents with each query, index terms are weighted according to their usage in reference to documents. The intuition behind this weighting metric is that each referring author is permitted one vote for each index term, those terms receiving the most votes are weighted highest, provided they are not frequently used in reference to other documents that they become useless as discriminators. 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. Queries to Rosetta retrieve the documents for which the query words are most heavily weighted.

Information seekers use Rosetta through a simple Web-based interface. Queries retrieve a list of the most appropriate documents according to Rosetta's retrieval metric. People submit queries to Rosetta using natural language; no Boolean operators or other query language features such as parentheses or quotation marks are necessary. I implemented no query language features in Rosetta because people rarely use such features [**55, 31**], therefore system performance is best measured in their absence. In response to queries, Rosetta gathers all documents indexed by the query words and sorts them based on the number of words 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 hyperlinking to a document, Web authors describe that document using language that is very similar to the type of language a searcher is likely to use in queries for the information that document contains. Therefore, in response to a query, Rosetta associates the most importance with documents that have been described by at least one referrer using all of the language contained in a query.

2.4. User Interface

Rosetta lists documents retrieved in response to a query with summary information that helps users quickly determine the reason why a document was retrieved and how it will be useful. See Figure 2.2 for a sample search results page. Each document is identified by its title and a list of authors. The title is hyperlinked so that when selected the user may download the document itself. As a summary, Rosetta displays a sample of the text written in reference to a document and containing words used in the query. These words are highlighted, as is the point at which the document was cited to make it easier for information seekers to process each summary. This is similar to the usage in ResearchIndex though Rosetta uses references as document summaries rather than merely as an additional piece of information for which a user can ask. Rosetta uses references as document summaries for the same reason it uses them to index documents - they concisely identify important information a document contains. Furthermore, in reading a research paper people often use the way in which an author cites other

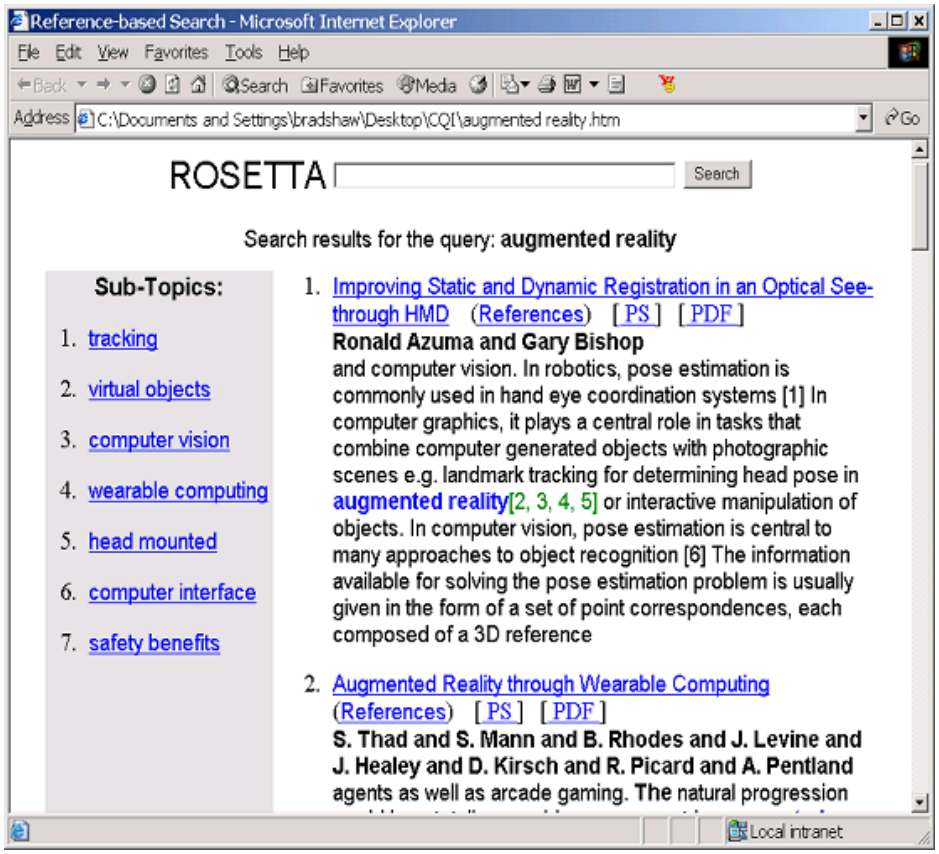


Figure 2.2. Rosetta retrieval results in response to the query "augmented reality". This example demonstrates that Rosetta identifies both alternate senses of the query such as "wearable computing" as well as important subtopics such as "tracking".

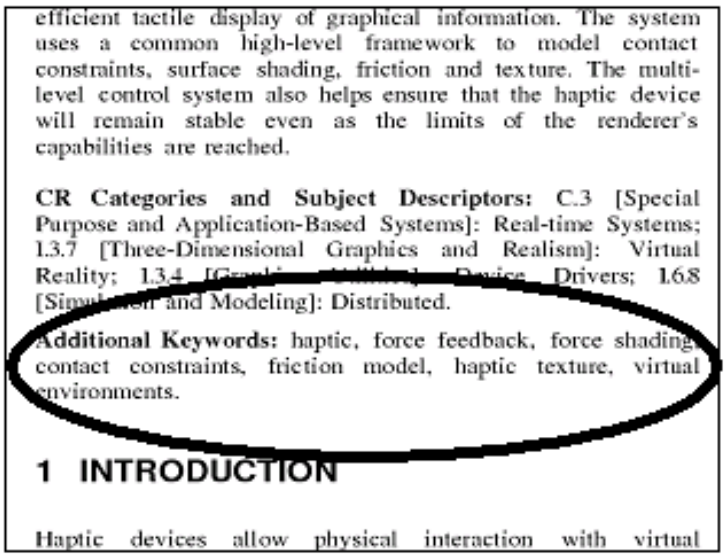
work to determine which bibliography entries will provide additional information of importance. Since people are already accustomed to using referential text to determine the value of documents, they will find it natural to use such text for the same purpose in Rosetta. In developing Rosetta, I implemented a generalized system that indexes documents using text written in reference to them. Using this

system I seek to demonstrate that referential text is better suited to the needs of a search engine than the content of the documents themselves. Primarily because, more so than content, the semantics driving reference are similar to the semantics of queries in that when referring to a document authors describe the information it provides, and when querying for information searchers describe the information they need. Furthermore, I seek to demonstrate that this is a general solution for collections of networked literature including hypertext. Rosetta merely uses text written to direct people to further reading without any processing specific to the domain of scientific literature. Using such data, Rosetta precisely identifies the value of documents. In later chapters, I successfully demonstrate that reference provides a better basis for indexing than content in collections of scientific literature. In addition, the results I present indicate that this result holds for other types of self-referencing literature such as hypertext.

CHAPTER 3

Search Performance

In this chapter, I explore Rosetta's performance as a search engine for scientific literature. In particular, I evaluate the precision with which Rosetta retrieves relevant documents and the significance of the contributions made by these documents. To demonstrate the advantages of Rosetta over traditional search solutions for collections of scientific literature I present this evaluation in a side-by-side comparison with a search engine implemented using traditional IR techniques. In particular this system is based on the Vector Space Model (VSM) [**47**] and employs standard TFIDF term weighting [**53**] and a Cosine retrieval metric [**50**]. For this set of experiments I evaluated Rosetta's retrieval performance on a randomly selected collection of 10,000 documents acquired from ResearchIndex [**36**] with the permission of Steve Lawrence. ResearchIndex collected the documents over which I ran this experiment from thousands of web sites at universities and other research institutions. They range in topic over many disciplines related to computing in one form or another, including Cognitive Psychology, Statistics, Computational Biology, and Computer Graphics among hundreds of others. Rosetta indexed these documents using the words found in text written in reference to them. The TFIDF/Cosine system in turn indexed them by the words used within the documents themselves.



efficient tactile display of graphical information. The system uses a common high-level framework to model contact constraints, surface shading, friction and texture. The multi-level control system also helps ensure that the haptic device will remain stable even as the limits of the renderer's capabilities are reached.

CR Categories and Subject Descriptors: C.3 [Special Purpose and Application-Based Systems]: Real-time Systems; I.3.7 [Three-Dimensional Graphics and Realism]: Virtual Reality; I.3.4 [Graphics Utilities]: Device Drivers; I.6.8 [Simulation and Modeling]: Distributed.

Additional Keywords: haptic, force feedback, force shading, contact constraints, friction model, haptic texture, virtual environments.

1 INTRODUCTION

Haptic devices allow physical interaction with virtual

Figure 3.1. Queries for the experiments described in this chapter were selected at random from keywords sections such as this.

3.1. A Study Using Contextualized Queries

I tested Rosetta's search performance on 25 queries with appropriate t-tests to determine the statistical significance of the results. As queries I used the words and phrases found in the "Keywords" section of 24 documents selected at random. By keywords I am not referring to the type of formal hierarchical subject descriptors employed by the Association for Computing Machinery (ACM) for their conferences and journals. Instead I mean the identifiers that researchers create for their work using their own words. For an example, see Figure 3.1.[1] I chose to test

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

Rosetta using queries collected in this way, because of a variety of problems inherent in other methods of evaluation. If I simply made Rosetta publicly available and asked people who are willing to submit evaluations, aside from the difficulties of attracting users to a new information system, I would likely only get feedback from people who really loved the system and those who hated it - not a particularly random sample. The TREC test collections for scientific literature (i.e. the CACM collection as used in [**47, 13**]) on which I might test Rosetta contain queries that are very different from the queries people naturally use when searching for information on-line. Many of the queries in these test sets are long sentences containing as many as twenty or thirty words, a far cry from the two or three word queries submitted by most users of search engines both on the Web [**55**] and in digital libraries [**31**]. Other methods in which test subjects use a system to look for any topic that interests them are artificial in that the searches occur in the absence of any real context. That is they are rarely prompted by a real information need, so any assessment of relevance is suspect. In contrast, in specifying a key word or phrase to describe his work an author uses his own words to describe a specific piece of information in which he is interested in a very real context – that of his research.

I selected the queries used in this evaluation at random from each of the 24 source documents. They varied in length from one to three words with 19 consisting of two words, three consisting of one word, and three composed of three words, for

adaptive estimation	sonic feedback
groupware	supervised learning
haptic	topology changes
hardware performance counter	transient interactions
inductive transfer	user interfaces
information sharing	virtual environments
reinforcement learning	virtual finger
reliable data distribution	visual reconstruction
reliable multicast	wait free
semistructured data	wavelets
shared variables	wireless routing
simulation acceleration	wrapper induction
software architecture diagrams	

Table 3.1. 25 queries used in experiments testing Rosetta's search performance.

an average of two words per query. See Table 3.1 for the complete list of queries used in this experiment.

For each query, in order to determine the relevance for a document retrieved I used the paper from which the query was drawn as the context in which that query was submitted. By this I mean that I used the source paper as a definition for what the query meant. For example, one query, "reliable data distribution", was drawn from a paper describing research on multicast technology for distributing bulk data such as video feeds to many clients simultaneously with error detection and congestion control.[2] For this query I marked as relevant documents that discuss multicast technology that ensures reliable distribution. For every query I used

[2]Query drawn from 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.

the paper from which a query was drawn as the context for the search and evaluated twenty search results (the top ten from both Rosetta and the TFIDF/Cosine system). I constructed a meta-search interface that searched both Rosetta and the TFIDF/Cosine system 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 it's origin.

3.2. A TFIDF/Cosine System for Performance Comparison

The TFIDF/Cosine system against which I measured Rosetta's search performance employs widely used and well-understood techniques that have proven to be among the best search technologies developed by the IR community [**49**]. It computes TFIDF term weight values, representing the significance of a particular word as a descriptor for a document using the expression:

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

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 [**47**]. The intuition here is that the best index terms for a document are those that are used frequently within that document and are fairly unique to it in that they are used in few other

documents. The TFIDF/Cosine system uses Salton's cosine metric to rank search results. This metric is described by the following expression:

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

This metric concisely described in [**50**] treats documents as vectors in an T-dimensional space, 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. Salton borrowed the cosine metric for document retrieval directly from vector algebra by mapping documents into a vector space [**47**]. In vector algebra, the cosine of an angle between two vectors is defined by:

$$cos(A,B) = \frac{A \cdot B}{\|A\|\|B\|}$$

in which the cosine of the angle between two vectors, A and B, is equal to their dot product divided by the product of their norms. Equation 3.2 is Equation 3.2 expressed in terms of document vectors. TFIDF/Cosine treats queries as document vectors in which all terms have a weight of 1. In response to a query it ranks each document based on the cosine of the angle between its vector and the query vector. In contrast, Rosetta favors documents that have often been referenced using the words in the query. Rosetta is implemented using term weighting and retrieval

metrics based on some of the same intuitions prompting Equations 3.2 and 3.2, but leveraging the power of repeated reference to documents in pinpointing the best index terms rather than the usage of terms within documents. For details on the implementation of Rosetta please see Chapter 2.

3.3. Retrieval Precision

Having evaluated the search results for each query I found that Rosetta compares very favorably to traditional IR techniques. In general Rosetta identifies documents relevant to queries with better precision, making fewer of the kind of retrieval errors common to standard vector-space techniques. Figure 3.2 depicts a side-by-side comparison over the 25 queries that comprise this experiment. While the two systems followed largely the same pattern of retrieval, reflecting variables such as query ambiguity and coverage of each topic within the collection, Rosetta exhibited greater retrieval precision for most queries. Figure 3.3 depicts the distribution of differences in the number of relevant documents retrieved. Rosetta performed better than TFIDF/Cosine for over two-thirds of the queries and as good or better for 80% of the queries. Out of the 5 queries for which TFIDF/Cosine performed better, for only 1 query was the difference in number relevant documents retrieved in the top ten greater than 1. Rosetta retrieved at least 3 more relevant documents than TFIDF/Cosine in the top ten for over half the queries and 4 more relevant documents or more for many of these. In contrast, TFIDF/Cosine retrieved 3 more relevant documents for only one query. On average Rosetta placed

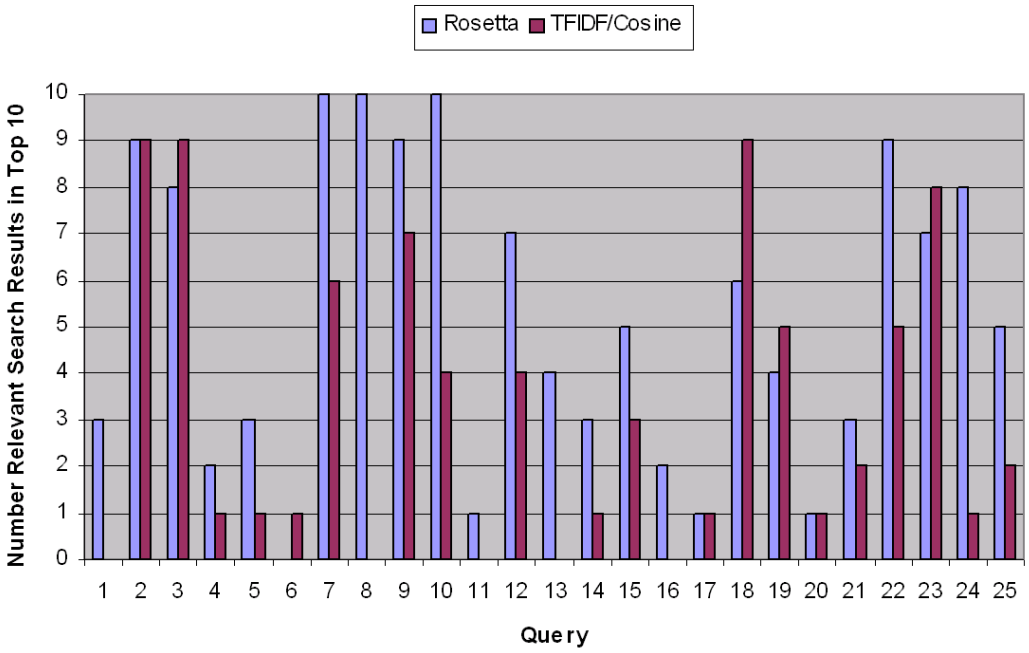


Figure 3.2. Number of relevant documents in top 10 search results. Rosetta's search performance compared to that of a traditional IR system using TFIDF for term weighting and the Cosine metric for ranking search results.

2 more relevant documents in the top ten than the standard IR system. A test of significance at a confidence level of 90% gives a margin of error of 0.9, indicating that RDI provides a significant boost in retrieval precision over traditional IR techniques.

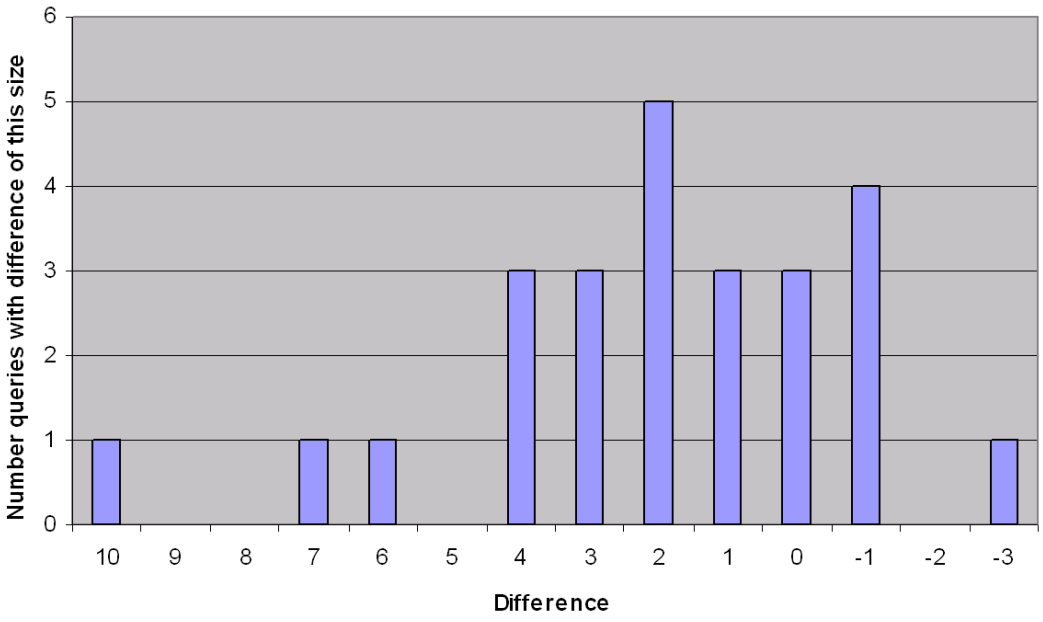
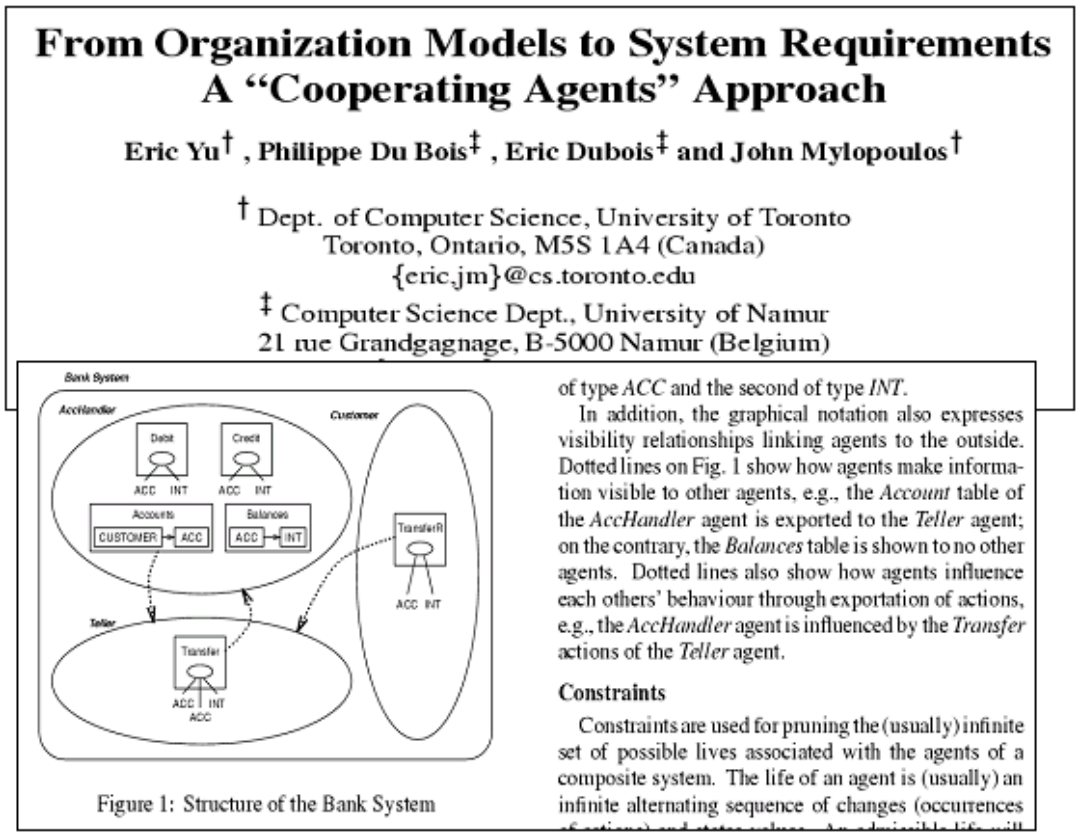


Figure 3.3. Difference in number of relevant search results in the top 10, comparing Rosetta and TFIDF/Cosine (Rosetta - TFIDF/Cosine).

3.4. The Problem of Relevance

People use most words in many different senses, even if a system chooses for each document only index words that in some sense identify the topic of that document, it may still incorrectly retrieve a document for a query assuming a sense of its index words other than that describing the information it contains. For example, the word "library" has one sense when used in the phrase "digital library" and another when used in the context of computer programming where it may refer to a library of code accessible to application developers. Setting aside problems

of word sense, in using statistical measures of word occurrence to identify good index terms, search system cannot really distinguish words that identify what a document is about from those that are simply used frequently. As a result, systems using such techniques often select index terms for documents that identify topics other than those the documents address. An RDI approach, on the other hand is less likely to improperly index documents and therefore, less likely to retrieve irrelevant documents than approaches in which documents are indexed by their content. This claim is further supported by a deeper analysis of indexing behavior in Chapter 4. Documents simply do not provide very good descriptions of themselves, especially when those descriptions must be compared to two and three word queries. Authors frequently include a great deal of information in their writing that has little or nothing to do with the main topics they address. They write with the goal of communicating a body of ideas with enough clarity so as to convey understanding and the in course of writing necessarily include information such as examples that often contain many words that might mislead an automatic indexing and retrieval system. For example, the paper represented in Figure 3.4 was retrieved by the TFIDF/Cosine system in response to the query, "inductive transfer", a topic in the Machine Learning area. Though this paper was retrieved with a cosine score only slightly less than a very relevant document (less than 0.02 difference), it has nothing to do with the topic in question. It was retrieved because the paper contains a very lengthy banking example in which the word



From Organization Models to System Requirements
A "Cooperating Agents" Approach

Eric Yu† , Philippe Du Bois‡ , Eric Dubois‡ and John Mylopoulos†

† Dept. of Computer Science, University of Toronto
Toronto, Ontario, M5S 1A4 (Canada)
{eric,jm}@cs.toronto.edu

‡ Computer Science Dept., University of Namur
21 rue Grandgagnage, B-5000 Namur (Belgium)

Figure 1: Structure of the Bank System

of type *ACC* and the second of type *INT*.

In addition, the graphical notation also expresses visibility relationships linking agents to the outside. Dotted lines on Fig. 1 show how agents make information visible to other agents, e.g., the *Account* table of the *AccHandler* agent is exported to the *Teller* agent; on the contrary, the *Balances* table is shown to no other agents. Dotted lines also show how agents influence each others' behaviour through exportation of actions, e.g., the *AccHandler* agent is influenced by the *Transfer* actions of the *Teller* agent.

Constraints

Constraints are used for pruning the (usually) infinite set of possible lives associated with the agents of a composite system. The life of an agent is (usually) an infinite alternating sequence of changes (occurrences

Figure 3.4. This paper contains a lengthy example that causes it to be retrieved erroneously for many queries having nothing to do with the topic of the paper.

"transfer" is used repeatedly. Similarly, many researchers make use of technologies, models, or other tools in their own work and as a result discuss these tools in their writing even though their own work is only indirectly related. For example, *Add a good example of this here* Other misleading index terms arise because some words are simply necessary in telling the story of a body of work, but do not really describe what a document is about. Often these words when used in other

senses identify information for which people are likely to submit queries. They are therefore a source of false positives in searches for those topics. For example, the paper, "Test Data Sets for Evaluating Data Visualization Techniques" by Bergeron et al.[3] was retrieved erroneously by the TFIDF/Cosine system in response to the query "reliable data distribution" is about creating test data sets for scientific visualization applications. It uses the word "data" frequently, but this word is so commonly used through the collection that its retrieval power is low for any document. However, because the authors discuss the appropriate distribution of values within the test data sets they create, "distribution" is a heavily weighted index term, and as a result the TFIDF/Cosine system ranked it number one in the list of search results for this query. Finally, in many papers authors repeatedly use several words that individually identify the subject of their research, but in combination identify altogether different topics. At retrieval time such documents are also a common source of false positives. For example, 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 not retrieve a single document directly related to this topic, while Rosetta found four. One paper placed in the top ten search results by the TFIDF/Cosine system, is entitled, "The Design of Mixed Hardware/Software

[3]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.

Systems".[4] This paper uses the words "architecture" and "software" several times and the word"diagram" repeatedly even though this word does not directly identify the topic of the paper. Other papers retrieved by TFIDF/Cosine are either about circuit diagrams or other software engineering topics and were retrieved because of similar overlaps in word usage. In contrast, Rosetta accurately identified a number of papers concerned directly with software architecture diagrams, placing three of the four it retrieved in the top five search results. This and other evidence provided by this study indicate that a powerful advantage in indexing by reference is that referrers rarely identify unimportant details concerning the documents they cite. Rather they describe what it is about a document that makes it useful and thereby indicate the queries for which a document should be retrieved. An additional benefit of a reference-based approach is that references classify a piece of information as addressing an important need of a research area even if the authors never do so for whatever reason. In some instances authors do not foresee the application of their work in a particular way. In others they document their research before the language describing the type of work it represents was formalized. Some work is simply adopted as important to a particular research community, though the author does not explicitly classify his work as part of that community. For example, in response to the query "wrapper induction", which identifies a body of work in which information extraction tools are automatically

[4]J. Adams and D. Thomas. The design of mixed hardware/software systems. In Proc. of the Design Automation Conference, 1996.

or semi-automatically derived from a set of example documents, Rosetta retrieved the paper, "Learning to Extract Text-Based Information from the World Wide Web" by Stephen Soderland.[5] Though Soderland never uses either of the query words in the body of his paper. (Two items in the bibliography of this paper do use at least one of these words in their titles.) Several references to this work; however, identify it as contributing to work in wrapper induction. As this example demonstrates, authors reference documents using language that indicates how they are used. The combined evidence from multiple references to a document can be exploited to direct information seekers interested in a particular topic to the documents that communities of experts in that topic have agreed are most useful. Reference captures not simply the ideas of an individual author, but the practical application of those ideas by many people who are in a position to judge their utility.

3.5. Utility of Search Results

Having measured the precision with which Rosetta retrieves relevant documents, I looked next to the significance of the contributions made by these documents in an effort to understand the overall utility of the information the system provides. I do not mean to suggest that the importance of a paper to a research community is the only factor affecting the degree to which it will be useful to information seekers, rather, it is simply one measure. However, this measure

[5]S. Soderland. Learning to extract text-based information from the world wide web. In Proc. of KDD-97, pages 251-254, 1997.

has been used throughout the history of IR work dealing with scientific literature [**58, 32, 26, 36**]. The significance of a research article plays a large role in determining the degree to which that document is useful, because it is the important contributions that shape an area of research. To add to such a body of research, one must be aware of the important contributions of other researchers following similar pursuits. Furthermore, while one can argue that citation frequency is not the only way by which one may measure the significance of a contribution, it is more difficult to argue that a document receiving many citations is unlikely to be useful to an information seeker. Practically speaking, for one reason or another it has been found useful by many other people interested in the same ideas.

To evaluate the significance of search results I measured the number of citations to documents deemed relevant to each query in the study. As this evaluation demonstrates, Rosetta finds documents that are useful for what they contribute to the body of knowledge identified by a query. Figure 3.5 contrasts the numbers of citations to documents retrieved by Rosetta and the TFIDF/Cosine system. It graphs the median number of citations per year for each document in the set of relevant documents retrieved by each system for each query. More specifically, I calculated the average number of citations for each 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 relevant documents retrieved by each system. I use the median instead of mean, because it is less sensitive to a single search result

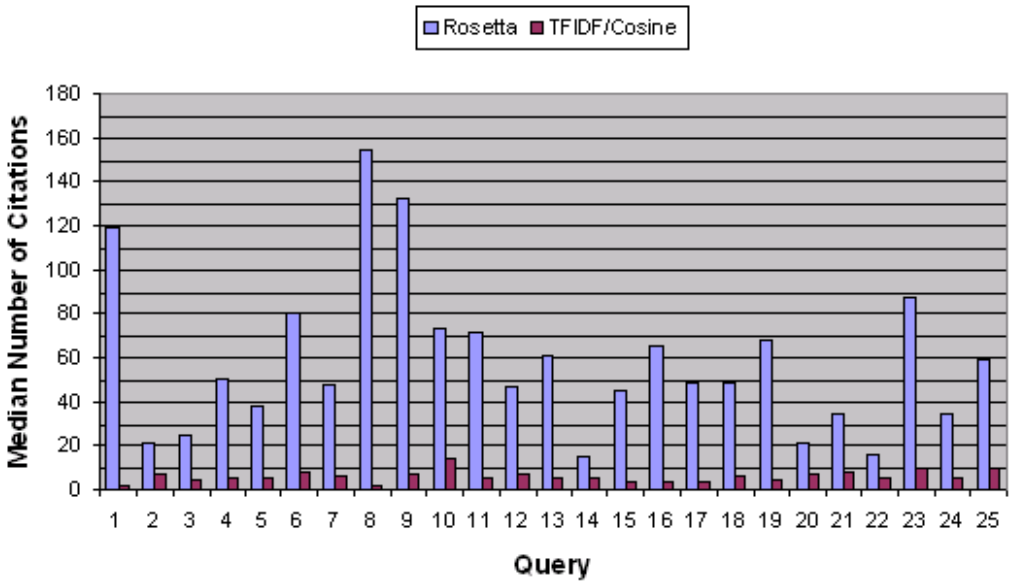


Figure 3.5. Median number of citations to relevant documents. A measure of the utility of the documents retrieved by Rosetta compared to those retrieved the TFIDF/Cosine system.

receiving many citations, and more reflective of the overall importance of the set of relevant documents in each set of search results. I 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 my 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, I measured the distribution of publication year for documents retrieved by Rosetta and the TFIDF/Cosine system. I found no significant difference between documents retrieved by the two systems. The mean year of publication for

documents retrieved by Rosetta is 1994, while the mean year of publication for documents retrieved by TFIDF/Cosine is 1995. Therefore, it is unlikely that the greater number of citations to the documents retrieved by Rosetta is simply the result of the age of the documents.

To be fair, standard IR techniques are not designed to rank search results on the basis of citation frequency so one would not expect to see a tendency toward high numbers of citations to documents retrieved in response to queries. Therefore, it is best to view the number of citations to documents retrieved by the TFIDF/Cosine system for a given query as a baseline for the number of citations to documents on that topic in the collection used in this study. As Figure 3.5 demonstrates, the number of citations to documents retrieved by Rosetta far exceeds the baseline. On average, the median number of citations/year to relevant documents for queries in this study was 52.44 greater than the baseline with a standard deviation of 36.12 and a margin of error of 11.9 at a confidence level of 90%. With all but the search results for two queries having a median of twenty citations or more and most having a median of at least forty, the results of this study indicate that Rosetta consistently retrieves documents that represent significant research contributions to the topic identified by a query.

The experiments presented in this chapter indicate that RDI achieves the goal of an IR technology for scientific literature that marries measures of relevance and utility to consistently provide search results that are useful for what they have to say in regard to a given inquiry. It is to be expected that a retrieval technique

closely tied to the number of citations a document receives will provide documents that are of greater research significance than techniques based on the frequency with which words are used within those documents. However, RDI as implemented in Rosetta appears to improve on the retrieval precision of such techniques as well. If Rosetta merely combined the retrieval of significant documents with a reasonable degree of precision I could argue that it provides more useful search results than systems based on traditional techniques, because for example, three relevant and significant documents may well be more useful than five relevant, but not particularly significant documents. However, Rosetta retrieves search results that are both more relevant and more significant that traditional techniques applied to the same data. The experimental evidence presented in this chapter strongly supports my claim that RDI provides much more useful information in response to queries than do traditional IR techniques when applied to collections of scientific literature.

CHAPTER 4

An Analysis of Indexing Vocabulary

In this chapter I explore at a moderate level of detail, the index terms Rosetta applies to documents in its collection. My objective is to demonstrate through a deeper analysis of index terms, the precision with which Rosetta indexes documents. In addition I present experimental that indicates that an RDI approach provides several other advantages over traditional indexing techniques. The most important of which include the selection of more diverse index terms, which helps to overcome the human-computer vocabulary problem as identified in [**24**] and the identification of meta-information such as whether a document is an introduction, overview, or presents information on a more specialized topic. As a means of comparison, I again make use of the TFIDF/Cosine system employed in Chapter 3.

4.1. The Study

I tested Rosetta's indexing performance on a collection of 10,000 research articles from a variety of fields of research. (The same collection used in Chapter 3. This collection represents a portion of those maintained by ResearchIndex [**36**]. As described in Chapter 2, Rosetta used windows of text surrounding citations to documents, windows approximately 100 words in length as the referential text

with which index those documents. I performed this study prior to that presented in Chapter 3. The version of Rosetta used in this study indexed documents using exactly the same term-weighting and retrieval metrics as the TFIDF/Cosine system described in Chapter 3. However, in later tests of search performance in which I compared this approach to the current implementation of Rosetta, I learned that the current implementation performs substantially better. Therefore, the study results presented here, though good, are likely not as good as they would be were the same experiments performed using the current implementation of Rosetta.

To perform these experiments, from the collection, I gathered a sample of 25 documents. One was later discarded, because it addressed a topic too far afield from my own, and was difficult to evaluate accurately. The documents selected were required to meet two restrictions, but were otherwise selected at random. First, I required that each document had been cited at least twenty times to reflect the lower bound on the median number of citations to documents in the top ten search results in queries to Rosetta. My goal in setting this restriction was to ensure that this experiment looked only at documents that information seekers are likely to review in lists of search results. Since studies of information seeking behavior indicate that many people rarely look beyond the first page of search results [**55, 31**], I choose the number of citations to documents considered in this study accordingly. Second, I required each document to contain a list of keywords specified by the author so that I could use these as indicators of what each document was about and therefore as a measure of the accuracy with

which the index terms extracted by Rosetta identify the important topics each document addressed. I imposed this restriction so that I did not introduce bias toward reference in determining myself which information should be identified by the index terms extracted for each document. As a point of clarification I should note that the keywords used in this study to identify the primary topics of each document were not the type of formal ontological subject descriptors employed by the Association for Computing Machinery (ACM) for purposes of determining reviewers for conference and journal articles. Rather they are the identifiers that researchers create for their work using their own words. As such they are more specific and do reflect the key ideas presented in the documents they describe at a finer level of granularity than subject ontologies.

Having identified the key features of each document included in the study, I then evaluated the degree to which each indexing vocabulary identified these features. Examining over 500 index terms for each document for both Rosetta and the TFIDF/Cosine system would simply be too time consuming to be practical, as a result I chose to evaluate only the top fifty index terms for each document extracted by each system. My reason for choosing fifty as the cutoff point was that based on a random sample of ten documents, the fiftieth most heavily weighted term serves as a good approximation of where for each system the weight of terms drops off significantly indicating that by both the metric employed by Rosetta and TFIDF, terms with a rank greater than 50 in this experiment do little to indicate what the document is about, and in more practical terms are unlikely to prompt a

high ranking for the document in search results, given that many other documents will almost certainly be more strongly associated with those terms.

4.2. Subject Precision

Documents are not self-summarizing; they are not intended to be. Rather an author's purpose in writing a document is to convey information with enough clarity to effectively communicate his ideas to his readers. For any topic this means that an author will use many words that make poor index terms with the same frequency and uniqueness as those that make good index terms. Such words occur in documents for many reasons. Some are simply words one tends to use when discussing a particular topic. For example, word usage in a press release from Handspring, Inc. describing their latest handheld may be such that the word "stylus" is heavily weighted as an index term, even though none of the new features described in the press release bear directly on the stylus or its use. Such a document might well be retrieved erroneously for queries from those seeking information on how to order a new stylus to replace the one they lost. Other words are weighted heavily as index terms even though they are not directly related to the topic of a document. Everyone understands more easily when given an example, as a result many authors include examples or draw analogies to topics with which their readers will be familiar in order to more effectively communicate by making new ideas easier to understand. In using such explanatory devices an author often introduce words that are not directly associated with her topic, but will fool

content-based IR techniques into weighting those words heavily either because they are used frequently in either long examples or because they are somewhat unique in the collection as a whole. Finally, for many documents some words weighted heavily as index terms are of questionable value, given that they are more likely to cause a document to be retrieved erroneously than they are to aid in the retrieval of a document when it should be retrieved. For example, a Computer Scientist in describing a distributed information gathering application written in the Java programming language may use the word "Java" throughout his paper and thereby cause that word to be weighted heavily as an index term. Since so many of these types of applications are written in Java it may or may not be a useful detail of the research presented in this paper. However, assuming the word "distributed" is also heavily weighted, such an assignment of term weights may prompt the retrieval of this document for distributed programming libraries written in Java such as that presented by [**22**]. Admittedly, though quite possible, this example is somewhat contrived, but I believe it illustrates the point that some details of a document named by some heavily weighted index terms should not be, because they may be more of a hindrance to retrieval performance than a help. While I do not address extensively such index terms in this study, near the end of this chapter I do present a small finding that indicates an RDI solution may more accurately pick and chose between those details that make useful distinctions and those that do not.

In measuring the precision with which Rosetta chooses index terms for the main topics of documents, I considered an average of 4.4 subjects per document

identified by the keywords found in those documents. Some of these topics included "shared variables" and "transient interactions" from a paper on mobile computing and "friction model" and "contact constraints" from a paper on a haptic (touch) interface for virtual environments. I compared Rosetta's performance to that of the TFIDF/Cosine system based on the degree to which each system weighted index terms that identified these topics more heavily that terms that identified other ideas. For each document, I marked as good terms, those used to name a topic identified by the keywords for that document as a main idea presented in the paper. By this I mean that in order to be considered a good index term, I required that at least one sentence actually used that term to identify a keyword topic in either the document itself or a piece of text used in reference to that document. More specifically, I considered good index terms to be those from any part of speech used to name one of the keyword topics either as an individual word or as part of a phrase. For example for the topic "contact constraints" the words "contact", "touch", and "touching" were all considered valid identifiers; and for a paper on Quality of Service both "quality" and "service" were considered good index terms. I performed no stemming for this study so in many cases multiple forms of a word appeared in the lists of indices I evaluated for a document. I found that on average only 34.9% of the content indices identified a topic also identified by one or more of the keywords for a document, while 50.5% of indices from reference identified the same subjects. Comparing the number of good index terms on a document-by-document basis, I found that the mean paired difference was 15.6%

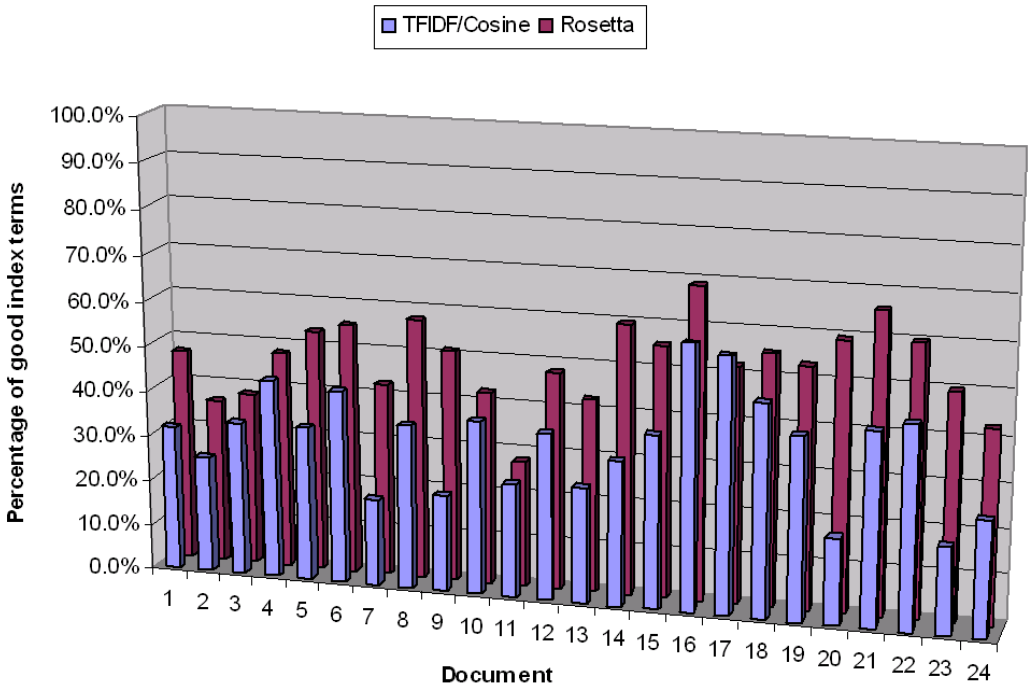


Figure 4.1. A comparison between Rosetta and the TFIDF/Cosine system on the basis of the percentage of the top 50 index terms for each document that accurately identify one or more of the primary ideas it addresses.

with a standard deviation of 10.3% and a 90% confidence interval of 3.5%. On average this means that out of the top 50 index terms for both Rosetta and the TFIDF/Cosine system, about 8 more of those selected by Rosetta identified a key subject for each document I evaluated. Figure 4.1 shows the relative precision of the indices drawn from content and reference and a percentage of the total number index terms considered for each document. The results presented here address only those topics for each document that the authors themselves identified as important

details in that they represent completely new ideas or topic areas in which the paper makes a contribution. I did not in this section address other topical details concerning the documents that may serve as important distinguishing features, but were not identified by the authors as such. In a later section in this chapter I do present a small study that explores index terms for potentially important topical features other than those identified by the keyword sections of the papers.

4.3. Index Terms Identifying Meta-Information

Most indexing and retrieval research ends with subject precision. However, an RDI approach, in addition to a greater ability to identify what documents are about, provides a second advantage in that such an approach identifies other features of documents that allow information seekers to easily distinguish among several documents on the same topic as to which will be more suited to there information needs. The type of distinguishing features to which I am referring are those that indicate the function of a document or what type of information the document contains, the type of knowledge that is often referred to as meta-information. For example, if an indexing system has correctly identified introductory texts on a given topic, then it can help information seekers with little or no background knowledge on that topic more effectively by either finding those texts in response to queries including the word "introduction" or by suggesting introductory material when available on the topic of an inquiry. Another type of information that would be of particular use to information seekers who wish to ramp up on a

subject would be documents that present an overview of that topic, because the organization components and high-level review is often a far more effective means of learning that beginning with the detail of individual documents In Computer Science as in most other topic areas the ability to make more specific distinctions between information germane to particular topic greatly reduces the amount of time searchers must spend skimming through a body of related information to find the type of information they need. For example, in developing new software technology a researcher will often find that some part of the problem with which she is working has already been well solved by others. In pursuit of her work, rather than building everything from scratch she will attempt to acquire as much supporting software as she can. While papers describing algorithms and theories relevant to the problems with which she is working are not without use, of more use would likely be papers describing software libraries or systems that are freely available for reuse by researchers such as her. An information system able to distinguish such papers will be of more use to this researcher than one that cannot. Following this reasoning I next tested Rosetta's ability to distinguish between different types of documents based solely on the index terms it extracted for them. In this phase of the study, I was interested in index terms that identified useful meta-information such as the type of contribution made by a document. As an example specific to this study, one document contained important study results, while another contributed a new algorithm in the area of mobile computing. To perform this component of the study I read the documents and determine what

meta-information, if any, was appropriately associated with them. I identified an average of 2 pieces of meta-information per document, but there were 2 documents for which I could determine no useful meta-information. The documents and corresponding meta-information for each is listed in Table fixme. As with subjects, I marked as good terms, words that named a piece of meta-information either singularly (i.e. "algorithm") or as part of a phrase (i.e. "study" for "study results"). I found that the index terms from the TFIDF/Cosine system identified the meta-information for a document in only 23% of the cases, while Rosetta index terms identified all meta-information for 50% of the documents I considered. Comparing the relative performance per document, Rosetta's index terms identified more meta-information for 64% of the documents and identified the same amount for 27% of the documents, leaving only 2 documents for which the TFIDF/Cosine system identified more meta-information. Figure 4.2 shows the relative performance of Rosetta and the TFIDF/Cosine system in identifying useful meta-information on a document-by-document basis.

4.4. Measuring Indexing Language Diversity

Finally, I wanted to get some indication of how well Rosetta might be able to handle queries for the same information coming from different searchers. As Furnas et al. point out in [**24**] different people use many different words to identify the same ideas. While in traditional IR systems, the content of documents identify many of these words, performing much better than a rigid manually constructed

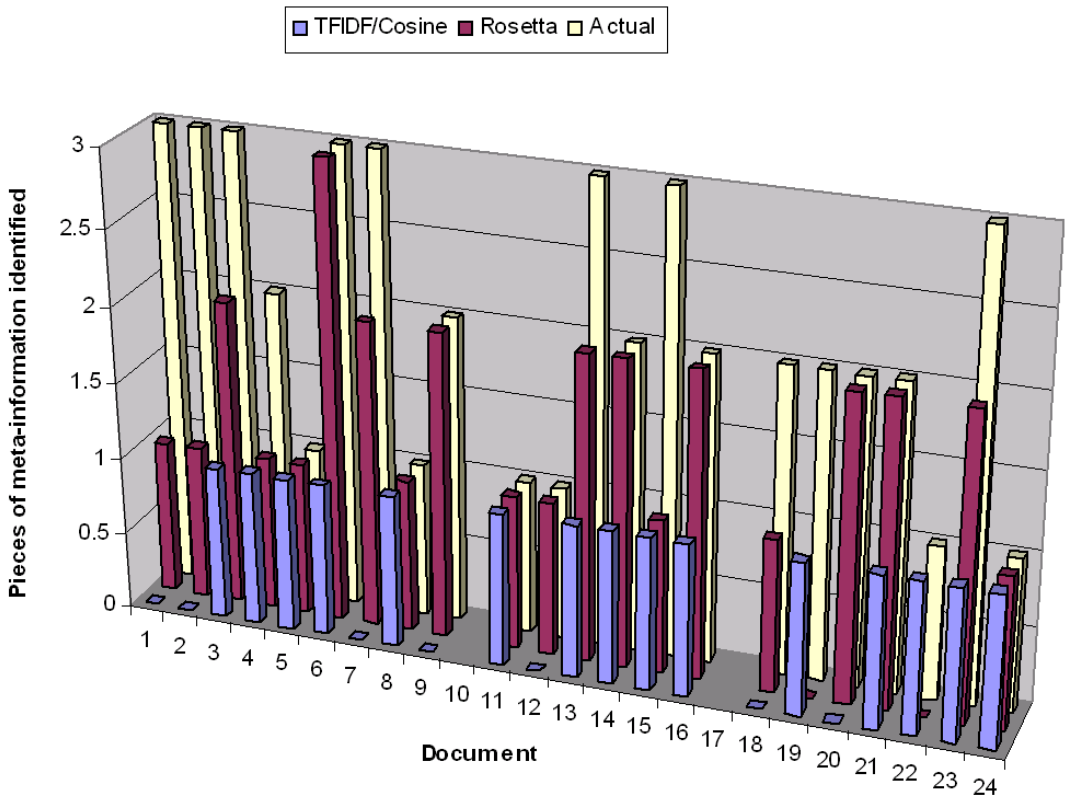


Figure 4.2. A comparison between Rosetta and the TFIDF/Cosine system on the basis of the number of pieces of meta-information accurately identified for each document.

ontology, I hypothesized that given the fact that referential texts are written by authors in a variety of research contexts, an RDI approach might identify more of the words people are likely to use in queries for the information documents contain, weighting heavily many terms that identify the same ideas for each document. Therefore, as the final phase of this study I measured the number of unique ways of identifying the information contained in each document represented in the index

terms extracted by Rosetta. As with the other phases of this study, I compared Rosetta's performance to that of the TFIDF/Cosine system. In this evaluation, I looked at the number of different ways in which any aspect of each document was described. I categorized the index terms for each document based on the way they were used in either the document or referential text. Essentially, I grouped together words used in phrases and multiple forms of the same word as single means of identifying some feature of each document. For example, for the topic of a haptic (touch) interface for a virtual environment discussed in one document, the indices "touch", "touching", and "interface" were grouped together as a single means of identifying that document. In addition, the words "haptic" and "display" were also grouped together as a second means of identifying this topic because the phrase "haptic display" appeared frequently in the content of the document. I captured the different groups of words an author strung together to identify some feature of a document and treated these as unique means of describing that document. As one further point of clarification, I did not count the number of aliases for a topic that could be formed using various combinations of the words that participate in at least one identifier for a concept. I only recognized unique identifiers that were actually constructed by either the author of a document or authors citing that document. For example, while the phrases "haptic interface" and "haptic display" were used to describe a document, the phrase "touch display" was not, so it was not counted as an additional unique identifier for a document. In evaluating the diversity of words extracted as index terms, I found that the average number of

unique identifiers per document identified by the TFIDF/Cosine system was 10.5 while Rosetta found 16.2 on average. The mean paired difference for each document was 5.7 with a standard deviation of 3.1 and a confidence interval of 1. Figure 4.3 charts the difference between the TFIDF/Cosine system and Rosetta as sources of unique identifiers for the documents. As I hypothesized, many authors citing a document in the context of their own work, do appear to bring out many different ways of describing the same idea. Each citation indicates a different perspective through the words used to describe the cited document. With typically twenty or thirty and as many as several hundred citations to valuable documents, the index terms extracted using an RDI approach create a larger target for searchers to hit than those extracted using traditional methods. In other words, queries arising from the particular context in which a searcher is working have a much better chance of matching the words used by many referrers than they do of matching only the words of an individual author.

4.5. Another Look at Subject Identifiers

Having reported Rosetta's performance when compared to the TFIDF/Cosine system against the three metrics of primary interest in this study, I return to the issue of topical precision. Combining the index terms for both topical and meta-information 52.8% of those extracted by Rosetta identified an important feature of some document in this study. This is compared to 35.8% of those extracted by the TFIDF/Cosine system. The question then remains – what did the other

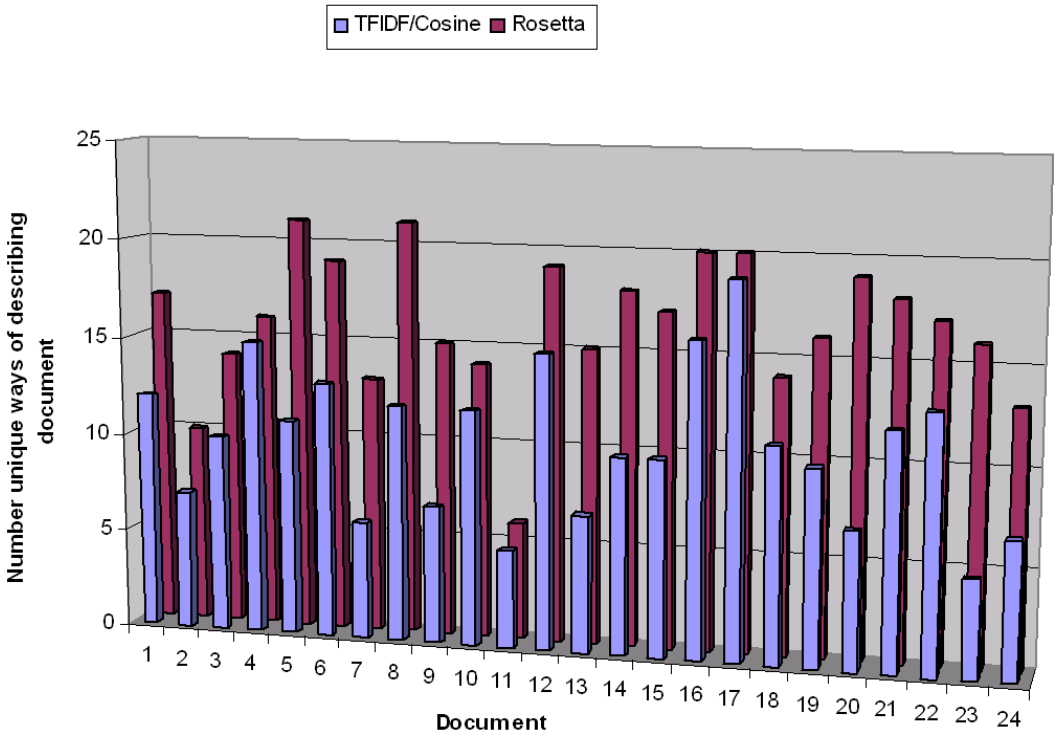


Figure 4.3. A comparison between Rosetta and the TFIDF/Cosine system on the basis of the number of unique terms each system identified as index terms for the set of ideas each document presents. The claim here is that a greater number of unique index terms will permit more people to find what they are looking for, because different people tend to search for the same information using many different queries.

indices identify? For both Rosetta and the TFIDF/Cosine system the overwhelming majority of these indices (approximately 70% for both sources) identified details about a paper in addition to the key topics and meta-information used in this study. They identified a wide variety of information including various details

about the implementation of a particular solution or the application domain in which a researcher works. In general, they identified various concepts an author addressed for one reason or another that were not central to the contribution of the paper in any way I was able to determine. However, that is not to say that these index terms identified no information that people would find useful. It is impossible to predict who will search for the information a document contains and what their motivation for such a search might be. Likewise, it is impossible to identify all the features, not to mention search terms, by which a document should be indexed. In this study, I originally identified the interesting features of a document by gaining an understanding of its contributions. To gather some idea as to whether or not the words identifying additional document features make good index terms I looked at the source from which the words were drawn. My goal was to determine why they appeared in the list of index for a document in both Rosetta and the TFIDF/Cosine system. I found that over 90% of the index terms extracted by Rosetta were weighted heavily because several researchers used that word to identify a document feature they found important. For the TFIDF/Cosine system such an evaluation is not possible, because no process of vetting indicates the features that make that document useful to other people. However, in an attempt to measure the degree to which the additional words extracted by the TFIDF/Cosine system serve as good index terms for a document I compared them to those extracted by Rosetta for a sample of ten documents used in this study. I found that the index terms extracted by the TFIDF/Cosine system identify only

63of the same features referrers considered important. While this result does not necessarily mean that the remaining index terms extracted by the TFIDF/Cosine system poorly identified what is useful about the documents I evaluated, it does mean that on average over one-third of these index terms identified features that not one of at least twenty independent reviewers (referrers) identified as important.

4.6. Discussion

The results of the study presented here indicate that the words authors of research papers use in reference to the documents they cite identify the subjects of those documents and other important features with precision, using a vocabulary that recognizes many different ways of describing the same idea. While by no means conclusive, these findings indicate that repeated citation of a document acts as a filtering process; identifying the important information a document contains in favor of other information that is not particularly interesting. In addition, authors who have cited a document serve as reviewers and recommend useful documents to the exclusion of those that are less useful for people interested in a particular subject. In performing the study I present here it was my goal to understand how well referential text might serve as the basis for an indexing and retrieval system for scientific literature. This study indicates that referential text precisely identifies most, if not all of the useful information a document contains with greater precision than the document itself and does so using a rich vocabulary. While this is by no means proof that an indexing technique such as the one suggested here

will provide retrieval performance superior to existing search technology, it does demonstrates that referential text better captures the essence of a document than the document itself.

CHAPTER 5

A Collaborative Query Interface

While the indexing techniques implemented in Rosetta show promise in improving the success with which search engine users find useful scientific literature using simple descriptions of their information needs, some information needs require additional user support. Given that most combinations of two or three words can be used in more than one sense, many queries will continue to be ambiguous no matter how precisely a system indexes the documents in its collection. In an attempt to overcome this problem, in Rosetta I have developed a Collaborative Query Interface (CQI) that suggests words to users that may help them more accurately specify the information in which they are interested. The motivation for this technology is that creating lists of words that unambiguously identify information needs is difficult. My goal is to transform this difficult generation task into a relatively simple recognition task. The idea here is to augment the set of search results retrieved in response to a query with terms that identify different senses in which the words of that query have meaning. For example, I recently did some work not related to this dissertation in which it was necessary for me to build some software that would semi-automatically build "wrappers" for web sites for use with a system that synthesized the results of queries to several information

sources (a meta-search system). By a wrapper I mean a tool accepts queries in a standard data language, submits them to the search site, and transforms them into a standard format recognized by the meta-search system. Knowing that research in this area is sometimes referred to as "wrapper induction" I submitted this phrase to Rosetta to see what results would come back from the collection used in the experiments in Chapters 3 and 4. Rosetta retrieved the set of results depicted in Figure 5.1. Based on the topic of the first paper, I realized that these two words are also used in a sense other than that I intended. A quick scan of the suggested set of query modifiers to the left of the search results indicates a "feature selection" sense in which these words are used together and an "information extraction" sense. Since "information extraction" was the sense in which I was interested, I selected the link identified by that label. The results of this new query ("wrapper induction" and "information extraction") are presented in Figure 5.2. As these search results demonstrate, with a single click I was able to effectively remove the ambiguity and locate a much more useful set of documents. In the remainder of this chapter I will first discuss the theory motivating this interface to Rosetta, I will then why an RDI approach is particularly well-suited to address this problem, I will then discuss the implementation of a CQI in Rosetta, and will conclude with some thoughts on work yet to be done with regard to this interface.

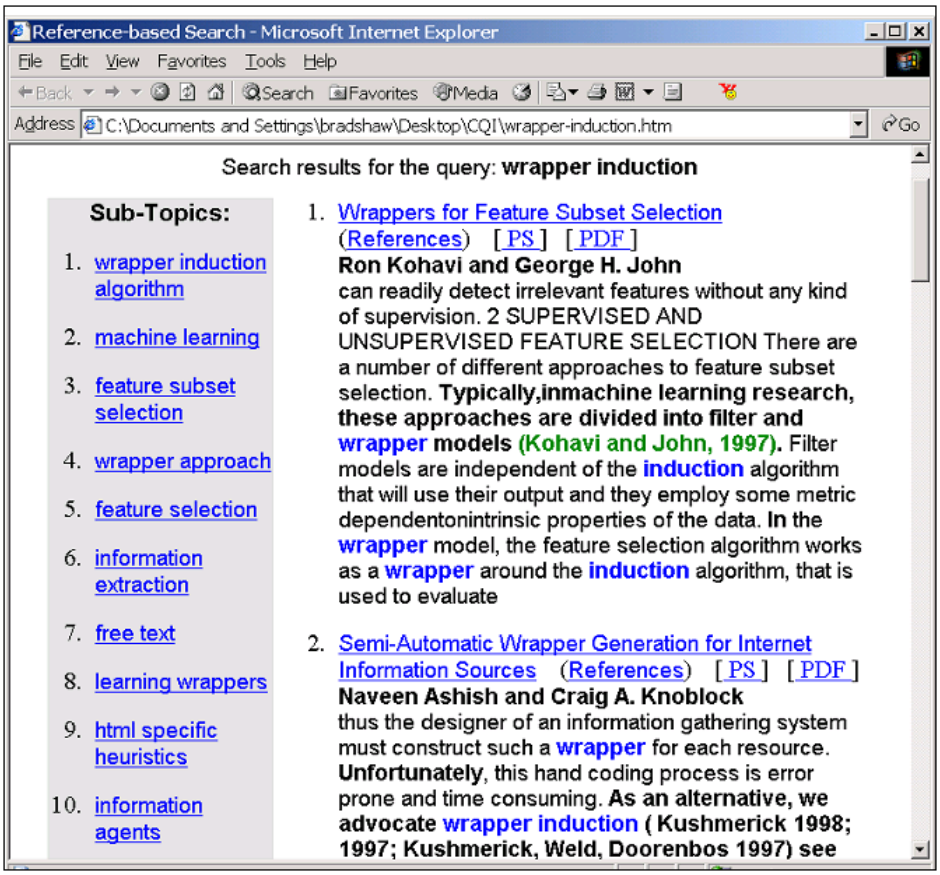


Figure 5.1. Rosetta's response to the query, "wrapper induction". Note the different senses of "feature extraction" and "information extraction" identified in the suggested list of query modifications or sub-topics.

5.1. Query Ambiguity: A Natural Consequence of Human Communication

The very nature of human communication presents an obstacle to creating unambiguous queries. Searchers usually do not work to create unambiguous queries because the same cognitive process that makes it possible for us to communicate

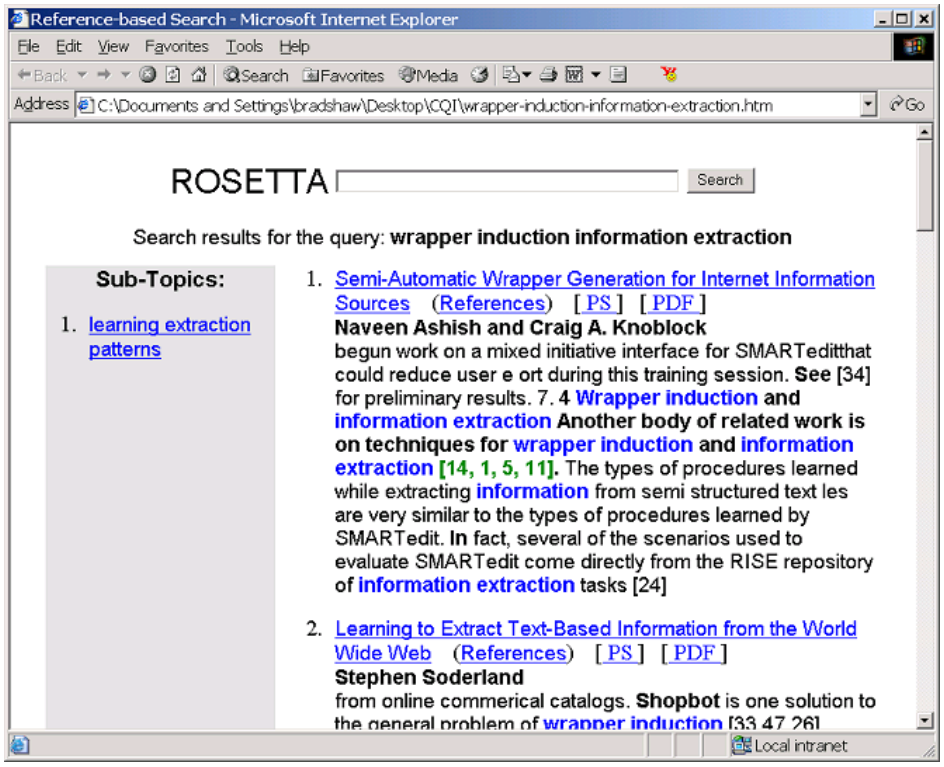


Figure 5.2. Rosetta's response to selecting "information extraction" from the set of query modifications suggested in response to the query, "wrapper induction".

effectively suppresses the other senses of terms they may use in queries. In the course of conversation people quickly work to establish context and from that point use that context to infer the appropriate sense of words and other semantics without even realizing it. This ability to suppress other meanings for the words someone uses permits conversation to continue in an efficient and effective manner. For example, we recently had some work done on our apartment building and during the construction the builders broke some of our windows. In a call to my wife

later in the day. I simply said, “The builders broke some of our windows today”. Because she and I shared the same context of work being done near our apartment there was no need to explain to what builders I was referring. I did not even need to further specify, the “kitchen windows”, because the context we shared was sufficient for her to infer all of the appropriate meanings. Imagine, if for every statement made in a conversation, you and the person with whom you were speaking had to go through a meta-conversation constantly resolving ambiguities arising over any conceivable interpretation of your words. Language would be a terribly inefficient tool; one wonders whether or not it would have ever developed. It is our natural ability and indeed, necessity to suppress all other meanings of the words we use in a specific context that make Web search an often difficult endeavor. Due to the context in which they are working searchers can rarely conceive of other senses of the words they use. As a result, initial queries are often terribly ambiguous. So why suggest additional query words, rather than simply letting users view the search results, realize the other contexts their query has identified, and make the necessary modifications to their queries in order to resolve the problem? The reason is that recognizing a disambiguating term is easier than generating one. People are adept at recognizing what they want when they see it, especially in search contexts. As a result, it is much easier to simply scan down a list of terms and select the one that identifies the context you intended than it is to recognize a common thread running through several irrelevant results and conjure a means of augmenting your query that will eliminate that interpretation of your query as a

possibility. Let alone one that will eliminate many of the other misinterpretations of your query. In short the cognitive process of recognition proceeds much faster than the analytical and generative processes one must undertake in order to modify a query effectively. Even expert searchers such as myself find the process of query munging in order to eliminate irrelevant results time-consuming and difficult. As a result, many users of information technology appear quick to abandon a search that does not immediately seem promising. In a study of the query logs from the Excite web search engine Spink et al. [**55**] found that users spend very little time on a single inquiry, most submitting only one or two queries for the same piece of information and looking at only a small percentage of the results. While this may simply indicate that people quickly find what they are looking for, given the general level of dissatisfaction people have with search engines, it is likely that this study demonstrates that people are either unwilling or unable to try a variety of approaches to find what they need, and are equally unwilling to wade through much irrelevant information to locate a few documents that will be useful to them. A study of users of the New Zealand Digital Library (NZDL) supports this interpretation. In this study, Jones et al. [**31**] looked at query sessions for users of the Computer Science collection (CSTR) in the NZDL. To view research articles in this collection, users must download the documents in either Postscript or PDF. Jones et al. report that the majority of CSTR sessions consist of two short queries and result in the download of 0 documents, a finding that causes the

authors to conclude that "...a substantial portion of users end a session without having met their information seeking goals."

5.2. A Proactive Approach to Resolving Query Ambiguity

Studies of search behavior support a more proactive approach to resolving query ambiguity, but the most important motivation driving the collaborative query approach is that ambiguity extends beyond word sense to simple under-specification. Though I described the "head mounted displays" example above as one of potential ambiguity with regard to sense if the search results for such a query are irrelevant to an information seeker the problem is more accurately described as insufficient specification of the information need. Imagine if Tom posed the following question to Mary: "Tell me about head mounted displays?" She would likely respond, "What about them?" Tom might say, "Can you tell me how well head mounted displays work for virtual reality applications?" Mary might still feel that the question is not specific enough and say so. Tom might then finally pose an answerable question such as "Do head mounted displays have sufficient resolution for virtual reality applications?" Stepping back into the world of search, this question might be submitted to a retrieval system as "head mounted displays resolution virtual reality"; however, such a query is unlikely. People rarely submit such long and specific queries to search engines. Spink et al. found that more than 86% of over one million queries to Excite contain three words or less [**55**], the average being closer to two words. In addition, this study indicates that most

people rarely use any query language features to make the intention of a query less ambiguous. She notes that people almost never put quotation marks around groups of words that they almost certainly intended to be phrases. Need to double-check this to make sure I'm citing the right article for each finding. Jones et al. [**31**] found that more technical users searched in much the same way. They report that over 80% of the 30,000 queries they reviewed contained three words or less. In addition, users of the CSTR appeared to take a similar approach to advanced query features as users of Excite. Few made use of Boolean operators to indicate the intersection (AND) or union (OR) of documents containing the words used in a query. Likewise few used parentheses or quotation marks to group words together in any way. In general people simply type two or three words and go. Many queries simply do not contain enough information to identify what the searcher needs with sufficient specificity. I believe the reason for this is that people unknowingly assume a shared context with information systems to the extent that they fail to even fully describe what they need let alone do what they can to resolve any confusion as to the sense in which their query words should be interpreted. Put a pop-culture example here with which people will identify. Therefore, searchers require tools to help them resolve ambiguity in their queries with regard to both problems of word-sense and under-specified information needs. In Rosetta I have developed technology based on RDI to help users describe their information need with sufficient specificity in an efficient and effective manner. An RDI approach is particular well-suited to this task, because RDI not only captures how documents are best described, but

it captures how people use words together to unambiguously identify each concept addressed in a collection of literature.

5.3. An Initial CQI Implementation

Rosetta uses a simple method to retrieve terms by which users might effectively extend their queries. In addition to documents, it also indexes the pieces of referential text from which index terms for documents are extracted. Each window of text extracted as a reference to some document in Rosetta's collection is indexed by the words used within it. In the currently implementation, at retrieval time, Rosetta selects a random sample of size s all references containing the query terms and compares each referential text to every other. It catalogs all two and three word phrases occurring in at least two references that pass a simple set of filtering heuristics. Having identified all phrases with which two or more references overlap, it then lists the phrases occurring in at least t references. In the examples presented above, s is set to 50 and t to 3. This process identifies phrases that commonly co-occur with all search terms.

Though I have not yet performed any experiments to test the utility of the query modifications this approach suggests, anecdotal evidence suggests it performs very well at identifying phrases that effectively identify more specific senses in which query terms are intended. However, anecdotal evidence does not permit me to make any claims about this approach. Instead, here I simply present the theory

behind the idea and an initial implementation. An empirical analysis of the CQI is planned as near-term future work.

CHAPTER 6

Conclusions

In this dissertation I have presented some quantitative assessments of the value of reference as an indexing tool, but have tried to do so in a way that points out the qualitative characteristics that make reference an important if not superior indexing technique for datasets for which it is possible to use such techniques. Though I have demonstrated specific metrics for term weighting and document retrieval, I believe the contribution of this work goes beyond the mathematics of the approach to a demonstration of the power in combining the evidence left by multiple users of a specific piece of information in determining the people for whom that information might be useful in the future. I am really talking about three ideas here. The first is that when authors reference a document for whatever reason, the language they use in the natural course of relating that document to their own work describes what that document is about. The second is that the sentences of different authors describing the same document, when compared reliably provide a consensus as to which words make the best descriptors and therefore, index terms for that document. The third is that the language many referrers use to describe a document is more likely to match the language that searchers use in queries for the same information. I have not proven the last statement in this dissertation,

but I believe it is important to argue for such a hypothesis here and will support this argument with some evidence from my work to date. When an author refers to another document he usually pinpoints the value of that document by naming the information it contains. Indeed, the very purpose of reference is to direct a reader to additional information that he may find useful, but to do so a referrer must describe it in such a way as to point out its significance and distinguish that from other information with which the reader may already be familiar. While referential text does not always include important index terms, any sufficiently useful document will be found, used and cited by people interested in the topics it addresses. For such documents, a sufficient number of references contain the words the best identify what that document is about and therefore, provide the data necessary to permit information retrieval systems to direct others interested in that information to the documents that will be most useful. The experimental evidence I have provided in this dissertation substantiates this claim. However, it is not individual references to documents that provide the true power of a reference-based approach. Rather it is the combined evidence of multiple references that establishes not only the significance of that information, but the topics about which it contains significant information.

6.1. Precise Indexing

Probably the most important contribution of the work I have presented in this dissertation has been in demonstrating the power of comparing multiple references

to a document as a tool for extracting index terms that precisely identify what that document is about. Stated simply, a community of people interested in the same ideas serve as reviewers for the information germane to those subjects and in their own writing begin to direct others to the information they have found useful in some way. In doing so, the community reaches a kind of implicit agreement not only as to which documents are important to the community, but also the reasons why those documents are important or useful. In identifying the use of one document or another, referrers naturally describe that document much of the same language. Terms used by many referrers to a document represent an implicit agreement by the community as to what the information in that document should be called. As I have demonstrated in earlier chapters, by determining the weight of a term on the basis of the number of distinct referrers that use that term in reference to a document correlates nicely with the degree to which it is a good index term for that document. While this may seem to imply that reference-based indexing identifies a small exclusive set of terms such those in topic hierarchies like the Library of Congress subject classification, such a situation would likely provide poor retrieval, because searchers would be require to come up with just the right set of words to find the information they needed.

6.2. A Broad Indexing Vocabulary

It was this problem that prompted the IR community to abandon topic taxonomies for indexing and retrieval decades ago and prompted some to suggest that

the content of documents be used as the source of words by which they should be indexed [**40**]. Furnas et al. [**24**] demonstrated that the content of documents provides fairly good coverage of the different words people might use to identify an idea. The experimental results I present in Chapter 4 indicate that a reference-based approach performs even better than the content of documents in identifying different ways people might describe and therefore, query for a particular piece of information. These results indicate that the words of many different authors typically reflect enough variation in the descriptions for a document that they identify many ways of naming each important idea presented in a document, more in fact, than the content of the average document. This result is intuitively pleasing, because it stands to reason that many people, working in several different contexts will use different words to identify the same idea. Combined with the results demonstrating the indexing and retrieval precision of Rosetta these results indicate that a reference-based approach not only chooses index terms for documents better than traditional approaches, but does using a diverse collection of terms that will match more users queries for the same information. Words used in reference to documents capture what those documents are about with far greater accuracy than do the words used within the documents themselves. The primary reason why this is true is that referential text is written to summarize a document, while the body of that document is written to convey information in detail. More precisely stated, referential text provides a better template against which to match

queries, because such text does not detail the ideas contained in a document; instead, it labels these ideas with words and phrases that people are likely to use in queries for the information they contain. As a result, the index terms chosen from referential text are less likely to contain words that people might regularly use to identify other important topics. Therefore, with an RDI solution it is less difficult to disambiguate the information of interest from other information around which the same query words may be used. The information seeking public, particularly those who are unfamiliar with the inner workings of the search technology they use are more prone to describe what they need using words and phrases that occur to them as natural descriptions of the information they need. For example, people searching the web for information on copying LPs to CD are likely to search using just these words. They are less likely to include disambiguating terms such as "audio input", "noise reduction", "hiss", and "pops" that will almost certainly be used in pages on this topic. Consequently, their search results consistently contain mostly irrelevant information such as those pages depicted in Figure 6.1 retrieved by Google in response to these queries. Text written in reference to documents tends to capture the way people naturally describe a piece of information, because they are often written by people who while writing are working in a context very similar to that of a searcher for the same piece of information. As a result, just a few words is often enough to supply the appropriate context, the correct "interpretation" of the query, and therefore, more accurate search results. The words "apartment" and "broken" are enough to identify the appropriate meaning of the

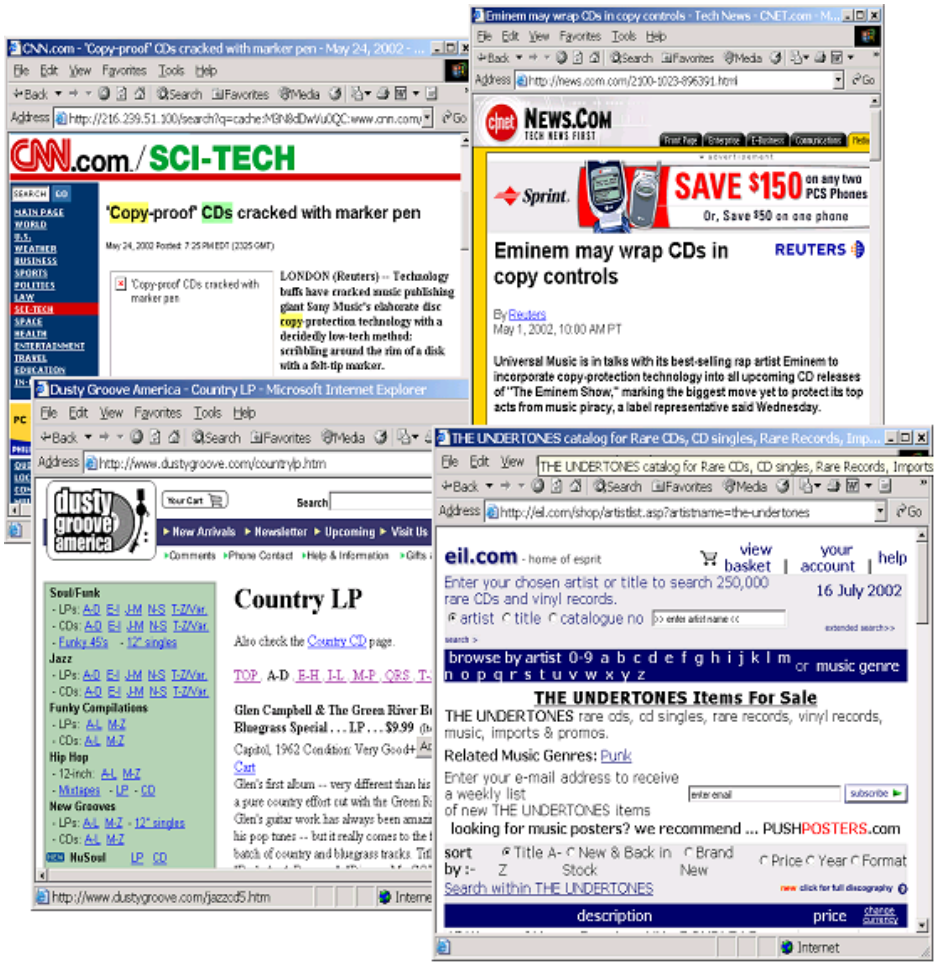


Figure 6.1. Search results for the query, "copy LP CD" demonstrating the ambiguity of such queries in systems where documents are indexed by the words used within them.

"window" so that the sentence, "The guys working on our apartment broke some windows.", was entirely comprehensible to my wife. Similarly, the sentence, "Use this software to copy LPs to CD.", unambiguously identifies the value of the object to which it is used in reference, whereas any number of documents might contain

the words “copy”, “LPs”, and “CD” for a variety of reasons, many of which have nothing to do with the topic of interest.

6.3. Relevance and Utility in a Single Measure

Greater accuracy in indexing and retrieval is one of the primary contributions of this research, but the true beauty of the RDI approach to building information systems is that it combines an accurate measure of what documents are about with an equally precise measure of their utility in a single, simple metric. Interesting documents attract the attention of many people, and the information age, this attention is easily tracked by the trails these people leave as they themselves create information. Based on the experimental evidence presented in this dissertation, the references to documents that people leave behind as they move through and contribute to an information space may be compared and contrasted to determine not only what terms most appropriately describe a document, but also how important that document is likely to be to people interested in information identified by those terms. By simply counting the number of people who use a particular term in reference to a document a system can automatically determine how well that term describes the document and how strongly people who have read that document recommend it to other people who search using that term. In addition, to the contributions I have made in demonstrating the superior nature of an RDI approach over traditional IR techniques, the contributions of this work extend beyond this to link-analysis work on the Web, particularly to that work

which has nibbled around the edges of the work I present here in using anchor text in one way or another as a source from which to draw index terms for the documents to which it links. To my knowledge, the enhanced topical precision of indexing and retrieval by reference has never been substantially demonstrated. More importantly, I believe my work is the first to demonstrate that in indexing by reference a system can regularly provide search results that are both on-point and significant in what they have to say about the topic of interest.

Though I have said throughout this dissertation that the combined evidence of multiple references to a document better identifies what a document is about than does the document itself, there is really something more subtle going on here. References do not identify all the information contained in a document; instead they identify the information that makes that document useful or distinctive. We have seen several examples of this with regard to the primary contributions of documents in scientific literature. In Chapter 3 I outlined several mistakes made by the TFIDF/Cosine system in its weighting of index terms, because of one type of information or another present in a document, but not germane to the primary contributions of that document. One example, was the document that contained a lengthy example using banking as a source of illustration, the example has little to do with the useful information in that document. The index terms Rosetta selected for this document reflect this; however in the experiments presented in Chapter 3 the TFIDF/Cosine system erroneously returned this document in response to a query as a direct result of heavily weighting index terms used in the example.

6.4. RDI Captures Meta-Information

But references go beyond simply identifying the useful ideas in a document and often include critiques, categorizations, or other assessments that are useful in distinguishing the utility of documents. In Chapter 4 I labeled this as meta-information about the value of documents. As an example, the following reference describes the paper by David Leake entitled CBR in Context: The Present and Future as one that presents a good overview of research in Artificial Intelligence in the area of Case-Based Reasoning.

> The basic idea of Case Based Reasoning (a good overview is given in [5]) is to solve new problems by comparing them to old problems that already have been solved in the past.[1]

In fact, more than one referrer to this document identifies it in this way though neither the title nor the words used in the document itself provide enough information to make this assessment using traditional IR approaches. Other types of assessments identify the function of the contribution an author makes. By function I mean the type of use to which information seekers might put the information gathered. An overview paper such as in the example above will likely have a very different function in the work of other researchers than does a paper that presents a new body of work in the same field. Other papers might present algorithms,

[1]From I. Vollrath, W. Wilke, R. Bergmann. Intelligent Electronic Catalogs for Sales Support - Introducing Case-Based Reasoning Techniques to On-Line Product Selection Applications. In R. Roy, T. Furuhashi, and P. K. Chawddhry (Eds.), Advances in Soft Computing - Engineering Design and Manufacturing. Springer. London, 1999.

theories, experimental results, or any of several other types of information that have greater or lesser value to information seekers depending on the task for which they need to acquire information. As a concrete example, in the paper pictured in Figure 6.2[2] the authors describe the Nexus message-passing library designed for use in parallel computing work in which complex calculations are performed by distributing the workload over many processors that communicate with one another from time to time in order to perform the task. This paper has been referenced several times using descriptions that identify it as presenting a message passing library. See the following for examples.

> This algorithm has been designed as part of NeXeme [25] a distributed implementation of Scheme, based on the message-passing library Nexus [10].[3]

> A Globus implementation [9] of this abstract communication device that uses the Nexus [10] communication library, and Globus mechanisms for resource allocation is available. [4]

> Various components of the Globus toolkit are described in detail in other papers [11, 7, 9, 13].[5]

[2]From I. Foster, C. Kesselman, and S. Tuecke. The Nexus Approach to Integrating Multithreading and Communication. Journal of Parallel and Distributed Computing, 37:70-82, 1996.
[3]From L. Moreau. Tree rerooting in distributed garbage collection: Implementation and performance evaluation. Higher-Order and Symbolic Computation, 14(4), 2002.
[4]From W. Benger, I. Foster, J. Novotny, E. Seidel, J. Shalf, W. Smith, and P. Walker. Numerical relativity in a distributed environment. In Proceedings of the Ninth SIAM Conference on Parallel Processing for Scientic Computing, March, 1999.
[5]From S. Brunett, K. Czajkowski, S. Fitzgerald, I. Foster, A. Johnson, C. Kesselman, J. Leigh, and S. Tuecke. Application experiences with the Globus toolkit. In HPDC7, pages 81-89, 1998.

The Nexus Approach to Integrating Multithreading and Communication

Ian Foster* Carl Kesselman† Steven Tuecke*

Abstract

Lightweight threads have an important role to play in parallel systems: they can be used to exploit shared-memory parallelism, to mask communication and I/O latencies, to implement remote memory access, and to support task-parallel and irregular applications. In this paper, we address the question of how to integrate threads and communication in high-performance distributed-memory systems. We propose an approach based on global pointer and remote service request mechanisms, and explain how these mechanisms support dynamic communication structures, asynchronous messaging, dynamic thread creation and destruction, and a global memory model via interprocessor references. We also explain how these mechanisms can be implemented in various environments. Our global pointer and remote service request mechanisms have been incorporated in a runtime system called Nexus that is used as a compiler target for parallel languages and as a substrate for higher-level communication libraries. We report the results of performance studies conducted using a Nexus implementation; these results indicate that Nexus mechanisms can be implemented efficiently on commodity hardware and software systems.

1 Introduction

Lightweight threads are often used to simplify the design and implementation of applications that must simulate or respond to events that can occur in an unpredictable sequence [32]. They are also, increasingly, used to exploit fine-grained concurrency in programs designed to run on shared-memory multiprocessors [19]. However, until recently lightweight threads were not used extensively in high-performance distributed-memory parallel computing, in large part because the focus of application development in this environment was on regular problems and algorithms. In these situations, the programmer can reasonably be assumed to have global knowledge of computation and communication patterns; this knowledge can then be exploited to develop programs in which a single thread per processor exchanges data by using send/receive communication libraries.

A number of factors are leading to an increased interest in multithreading in parallel computing environments, notably more irregular, heterogeneous, and task-parallel applications,

*Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439.
†Beckman Institute, California Institute of Technology, Pasadena, CA 91125.

1

Figure 6.2. Example of a paper presenting a contribution with a very specific function. The contribution is a message-passing library for parallel computing.

Globus uses Nexus [59] as its underlying communications tool.[6]

[6]From Katrina E. Falkner. The Provision of Relocation Transparency through a Formalised Naming System in a Distributed Mobile Object System. PhD thesis, Department of Computer Science, The University of Adelaide, May 2000.

> In [6], Foster et al. describe the problems involved in implementing Nexus' message delivery mechanism on various operating systems and hardware.[7]
>
> Nexus [7], a library for building distributed systems, has two salient features: a remote service request is a form of remote procedure call [1] and global pointers provide for global naming in a distributed environment.[8]
>
> We intend to use multithreaded point-to-point message passing packages such as Nexus [14][9]

Each of the references describes the paper as presenting a library that can be used for message passing in parallel programming applications. This distinguishes it from other work dealing with less complete solutions to message passing problems and in so doing identifies Nexus as a package people interested in building distributed applications would find particularly useful. Note that not only do these references identify the Nexus paper as one describe a message passing library, but they do so using a diverse set of terms that identify this topic - terms that are likely to match the queries of several different people interested in the same information. These references use various terms such as "message-passing", "communication",

[7]From K. Langendoen, J. Romein, R. Bhoedjang, and H. Bal. Integrating polling, interrupts, and thread management. In Proceedings of Frontiers'96, pages 13-22, 1996.
[8]From L. Moreau, D. DeRoure, and I. Foster. NeXeme: a Distributed Scheme Based on Nexus. In Proceedings of the Third International Europar Conference, pages 581-590, August 1997. Lecture Notes in Computer Science, volume 1300.
[9]From S. Hariri. ATM high performance computing laboratory. Syracuse University. 1997. (Unpublished document)

and “message delivery” to identify the central topic of the paper. Another represents a more general view of this idea, describing the paper as presenting a “library for building distributed systems”. Then, to describe the function the contribution of this paper serves, referrers use a variety of terms including “library”, “tool”, “toolkit”, and “package”. So the references to this paper hit the main topic of “message passing”, then identify the specific contribution in this area, that being a library for message passing, and finally do so using language that will match the queries of many information seekers working in a diversity of contexts. The point here is not that the content of documents presenting these types of contributions do not contain the words that identify such contributions. Rather, my claim is that because they are contributions that are intended to be reused in various ways by other researchers, many authors discuss such uses in their own work, so that it is nearly impossible based on the content of a document to distinguish the use of a particular technology from the presentation of that technology as a body of work in and of itself. Therefore, even when associated with specializing terms that identify a particular topic traditional IR systems often result in so many irrelevant documents that few searchers even think to specifically request them, and those that do often fail to find what they are looking for. For example, the query “knowledge representation system”, identifying a need for a general-purpose tool for use in work on knowledge-based decision making systems, will almost prompt the retrieval of documents literally all over the map Artificial Intelligence (AI) research. However, a reference-based approach will not only overcome the ambiguity

of one of the most generally used terms in nearly any collection, "system", but actually use that term to refine the search results to provide not only documents on "knowledge representation", but systems built and used by people for this purpose. For example, two of the top search results retrieved by Rosetta in response to this query are identified by the following two references which exemplify the type of descriptions found in references to each of the documents:

> The knowledge representation system used in the implementation is Classic [Borgida et al. 1989] a well-known, general-purpose knowledge representation system.[10]

> The Information Manifold [16, 15] is a system for building a knowledge base representing the user's interests.[11]

Because it provides this kind of specificity in indexing a reference-based approach successfully handles queries that would be nearly impossible to deal with using a traditional content-based technique. This is further demonstrated in other distinguishing features references attribute to all kinds of information. For example, returning to the Nexus paper above, though the paper itself never uses the word, the system is portable. That is it may be used on a variety of computing platforms, but such a summary is very difficult to extract from the content of the document,

[10]From G. De Giacomo, L. Iocchi, D. Nardi, and R. Rosati. A theory and implementation of cognitive mobile robots. Journal of Logic and Computation, 9(5):759-785, 1999.
[11]From S. W. Loke, A. Davison, and L. Sterling. Lightweight deductive databases on the world-wide web. In Proceedings of the First Workshop on Logic Programming Tools for INTERNET Applications, September 1996.

because rather than using such summarizing words as "portable", the authors instead talk about testing it on multiple platforms. References to this document, on the other hand, do summarize it as portable. The following reference is one example:

> We seek to address the requirements outlined in the preceding section by constructing a secure communications infrastructure based on a portable communications library called Nexus [12].[12]

For other documents rather than functional documents, refers list attributes either good or bad of particular solutions to the problems addressed in a paper. For example in the following reference:

> Active Message[8] is a fast message handling scheme with the address of the message handler contained in the message header.[13]

the referrer describes the cited paper as presenting a message-handling scheme with one major contribution being a technique that is extremely efficient or fast. Others demonstrate even more specific assessments of documents often involving critiques of one form or another such as the following:

---

[12]From I. Foster, N. Karonis, C. Kesselman, G. Koenig, and S. Tuecke. A secure communications infrastructure for high-performance distributed computing. In Proceedings of the Sixth IEEE Symposium on High-Performance Distributed Computing, 1997.

[13]From Junghwan Kim, Sangyong Han, Heunghwan Kim, and Seungho Cho. A New Communication and Computation Overlapping Model with Loop Sub-Partitioning and Dynamic Scheduling. In Proceedings of the ISCA Twelfth International Conference on Parallel and Distributed Computing Systems, 1999.

> Sreedhar and Gao described another approach which traversed the dominator tree of the program to compute J sets on demand [SG95] This algorithm requires O(E) preprocessing time, preprocessing space and query time, and it is easy to implement.[14]

In which the referrer identifies a particular algorithm for compiler program analysis as being easy to implement, a feature that might be of particular interest to someone with the need to rapidly develop a piece of technology. In other situations referrers critique an author's treatment of his subject, as is the case in:

> A complete discussion of the subject can be found in [3] while [11] and especially [15] give a detailed description of erasure codes more closely related to their implementation.[15]

where the referrer describes the cited work as a detailed discussion/description of erasure codes, a technique for error correction in multicast data distribution research.

As these examples illustrate, the power of a reference-based approach to IR extends beyond precise identification of the primary topics of documents to the identification of additional discriminating features, and finally to critiques of the documents themselves, all of which make it possible to separate useful information from that which is not with far greater specificity and precision than in traditional

[14]From Gianfranco Bilardi and Keshav Pingali. The static single assignment form and its computation. Cornell University Technical Report, July, 1999.
[15]From Luigi Rizzo and Lorenzo Vicisano, RMDP: An FEC-based reliable multicast protocol for wireless environments. Mobile Computing and Communications Review, 2(2), April 1998.

approaches. Such assessments make it possible for information seekers to find information that is not merely relevant in some way to a query, but has been truly distinguished for the reasons they specify in their queries.

6.5. Finding Good Examples of Bad Ideas

The true advantage here is that this technique enables you to find information that is distinguished for any reason, many of which are simply not identifiable from the content of a document. Imagine, for example, that you are actually interested in documents that are of poor quality. For this example, I will step outside the research paper domain, because academic culture is such that researchers rarely criticize other work with strong derogatory statements in their writing. Such civility is rarely in evidence on the Web; however. For example, for an undergraduate class I teach on Web technology I typically want to provide the students with examples of sites that are poorly organized, difficult to use, or in some other way, badly designed. A real web site will never describe itself as being an example of poor design; however, others will do so with glee. There are thousands of people interested in good design and many of these are prolific on the subject. In their writing they identify many examples of poor design. For example Figure 6.3 shows a note from Giles Turnbull, a regular poster to WriteTheWeb.com, describing time4.net as an example of what not to do when designing a web site. He writes:

> A new UK-based portal, Time4.net, commits several horrible crimes of bad web design. But the people who built it are aiming

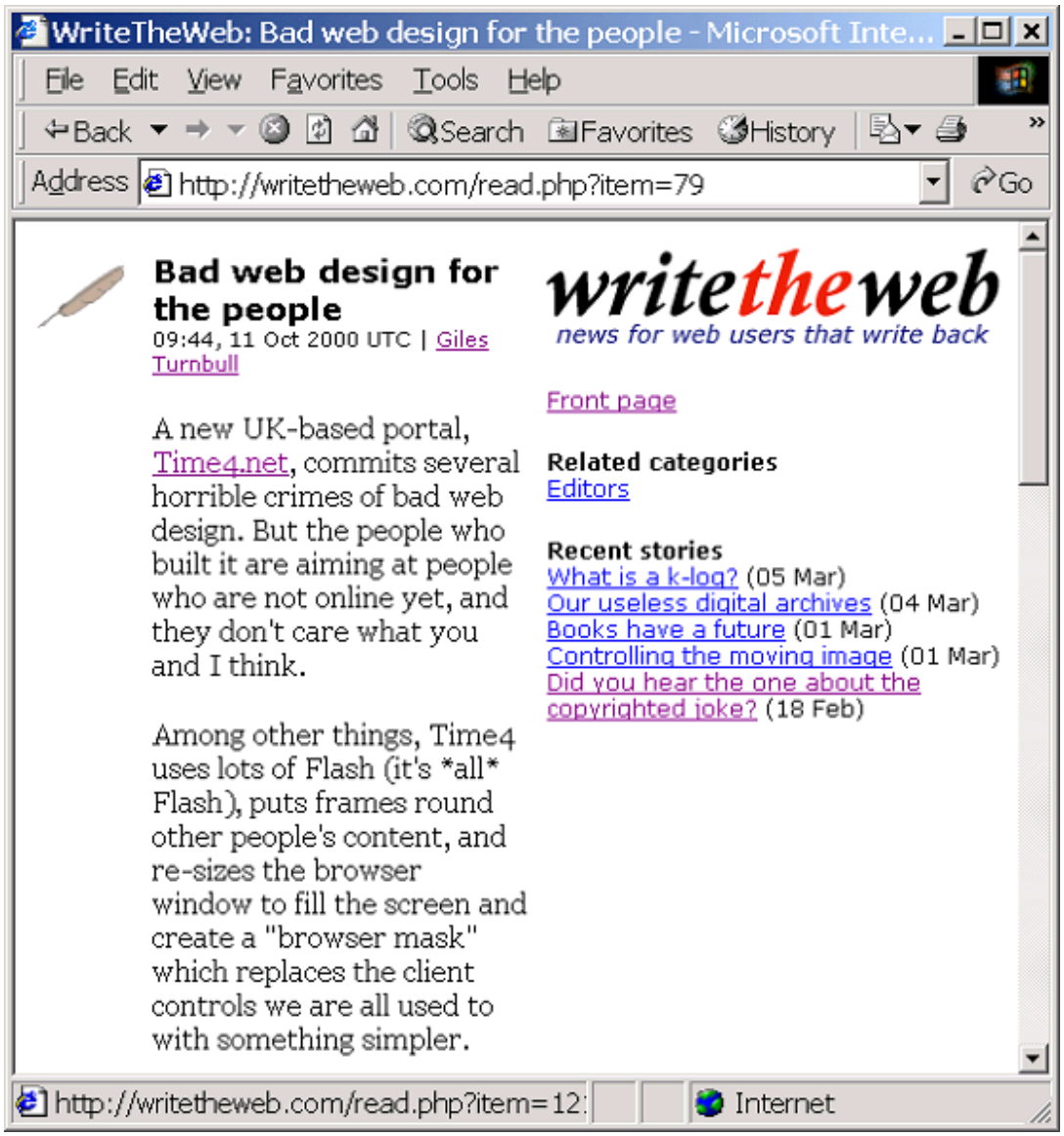


Figure 6.3. An example of a reference to a Web site that is useful not because it is good, but because it is bad.

at people who are not online yet, and they don't care what you and I think. Among other things, Time4 uses lots of Flash (it's

> *all* Flash), puts frames round other people's content, and resizes the browser window to fill the screen and create a "browser mask" which replaces the client controls we are all used to with something simpler.

Mr. Turnbull describes this page using the phrase "bad web design" - quite possibly the most likely query for such an example. He goes on to describe why it is bad, highlighting problems such as too much Flash, window re-sizing, and removal of the standard browser controls. In the process he has not only created a reference that will match queries for examples of bad web design, but also those for examples of specific design errors commonly encountered such as the over-use of Flash and various control issues involving the browser. In referencing other pieces of information authors describe what makes them distinctive, and why you as their reader should have a look for yourself. Repeated reference to a document not only strengthens ties to its distinctive features, but also helps to filter out potentially misleading identifiers. In the process people specify not simply which documents are good or even which documents are good for a particular reason, rather they describe the features that make a document distinct from those that are similar and specify various reasons why you might be interested in that document, whatever those reasons might be.

Finally, researchers have worked on using collocation information for years in IR. A document that is cited regularly serves as a focal point for collecting co-occurrence information. This is the theory supporting the development of the

collaborative query interface in Rosetta. But this also has a great deal of value in making a reference-based approach to indexing simple and straightforward, but more powerful than content-based approaches. For example, Rosetta rarely makes a mistake in failing to identify words intended as phrases to be words used as phrases. Conversely it does not require an interface that assumes phrases and in so doing makes the mistake of either not retrieving documents that should be retrieved, but don't happen to use the particular phrasal construction chosen by a searcher or of assuming phrases were found when words were simply used near one another.

6.6. RDI Ignores Large Volumes of Noise

Beyond the primary contributions of this research, the experimental evidence I have provided here demonstrates other benefits to the RDI approach. These results indicate that the technique is also quite robust. That is, it is able to deal with a large volume of noise in the text from which it extracts index terms. Again, this is because it uses the combined evidence of multiple references to documents to determine the weight of index terms. This is an absolutely essential feature of this technique if it to be all generally applicable, because the amount of text surrounding any point of reference that actually talks about the cited document varies significantly from reference to reference. For example, the following piece of text was used by Rosetta to index the paper, "The office of the future: A unified

approach to image-based modeling and spatially immersive displays." by Fuchs et al.:

> The Office of the Future project at the University of North Carolina [20] envisages an office adorned with a multitude of inexpensive cameras and projectors that are used to infer the geometry and reflective properties of all surfaces.[16]

Most of this reference talks about the document in question. In contrast, the following piece of text was also extracted as a reference to this paper.

> The display set up has been designed by Henry Fuchs[8].[17]

This reference contains very little in the way of good index terms. Most of the words that serve as good identifiers for what this document is about occur in sentences prior to this snippet. In scientific literature, at least, other problems can arise as well, in which the text selected actually contains descriptions of other documents, as in the following reference where several documents are cited in a list, with brief labels for each of them occurring just before the point of citation.

> These include the Power Wall at University of Minnesota [29] the Office of the Future project at University of North Carolina [22], the Interactive Workspaces Project at Stanford University [14]

[16]From Mark Ashdown and Peter Robinson. The Escritoire: A personal projected display for interacting with documents. Technical Report UCAM-CL-TR-538. Computing Laboratory, University of Cambridge, England. 2002.

[17]From J. Mulligan and K. Daniilidis. View-independent scene acquisition for tele-presence. In Proceedings of the International Symposium on Augmented Reality, pages 105-108, 2000.

> and display wall projects in various national laboratories such as Argonne [12] Lawrence Livermore [24] Sandia [10] and National Center for Supercomputing Applications [20] etc.[18]

This is problem is complicated by situations such as those demonstrated by the text below in which, a brief description that applies to all the documents is cited that contains useful index terms that apply to all of the documents is used at the beginning of a sentence in which, again, several documents are cited in a list, each with their own distinguishing labels.

> Image based techniques have also been used to extract multi layer 3D representations from 2D photographs [51] and to design the office of the future [83].[19]

While I have successfully written parsers that extract approximately 70% of the text in situations such as these that refers to documents other than the one in question, such parsing does not appear necessary to achieve good performance with this technique. In each of the experiments presented here, references to documents were used as is. While I am certain that retrieval performance could be improved to some extent by preprocessing references in an attempt to remove words used in reference to documents other than the document of interest, experimental evidence suggests that comparing and contrasting the words of multiple referrers

[18]From Han Chen, Kai Li, Thomas Funkhouser, Grant Wallace, Perry Cook, Anoop Gupta. Experiences with Scalability of Display Walls. In Proceedings of the Seventh Annual Immersive Projection Technology Symposium, 2002.

[19]From Emilio Camahort. 4D Light-field modeling and rendering . PhD thesis, Department of Computer Sciences, The University of Texas at Austin, May 2001.

significantly reduces the adverse affects of words not written to describe a document, but mistakenly captured within windows used as a references with which to index documents. Because indexing systems need only roughly estimate the text with which authors have referred to a particular document, the reference-based approach described in this dissertation might well be broadly applicable to all document collections in which windows of text surrounding links to other documents are easily identified and extracted. The most notable of this type of collection would, of course, be the Web.

6.7. RDI Improves As The Collection Grows

A reference-based approach is not only designed to deal with a lot of noise in the input text, but in contrast to traditional approaches, it may very well improve in search performance as more documents are added to the system. This claim can only be truly substantiated through future experiments, so here I will instead argue that it simply scales to larger and larger collections of documents better than traditional approaches. Traditional methods base the weight of an index term on the frequency with which that term is used in a document, they, in almost every case, weight heavily many terms that make poor index terms. The reason for this is that an author in telling his story cannot help but use frequently a variety of terms that are poor identifiers of what that document is about. After all the purpose of a document is not to summarize and label the information it contains, but to convey that information to its readers. As a result, indexing approaches

based on content are prone to retrieval errors, for the simple reason that there are countless reasons why a term might be used frequently within a document other than its use as an identifier for some important idea that document addresses. Therefore, the addition of a single document with heavily weighted index terms that do not identify what that document is about, but as is the case with nearly all terms, do identify other topics, search results for those topics will be adversely affected. This is the reason why so many search results for Web search engines are off-point. There are many Web pages with heavily weighted index terms that do nothing to identify what they are about, and as a result find their way into queries for wholly unrelated information. In a reference-based approach, on the other hand, in order for a document to be retrieved in response to queries for which it is not relevant, the words of a significant percentage of the referrers to a document would need to include the query terms, because the weight of an index term for a document is directly related to the number references to that document in which it is found. As a result, even if references were as prone to contain poor index terms as the content of documents, it would take the addition of several documents to significantly degrade the retrieval performance for any one topic.

6.8. RDI is Easy to Understand and Implement

Finally, in looking at the algorithmic benefits of this reference-based approach I should not fail to explicitly direct the reader to consider its simplicity. The technique requires no preprocessing, phrases recognition or other co-occurrence measures to perform well. It extracts single words as index terms and weights them using a simple TFIDF-like metric. In addition, the retrieval metric uses a simple weighted Boolean technique with which documents referenced using all query terms rank highest in search results, followed by those referenced using some subset of the query terms. The success of this method validates a simple and powerful method for finding documents that are well described by an entire query. In contrast, content-based approaches often rank some documents highly merely because some query words are strongly associated with a document, while other query words that serve an important distinguishing function are not represented at all. For example, at least two queries in the study on search performance presented in Chapter 3 contain important discriminating terms that the TFIDF/Cosine system failed to account for effectively in its search results. These queries were "reliable multicast" and "software architecture diagrams". For both of these queries Rosetta identified at least three more relevant documents in the top 10 than did TFIDF/Cosine. The problem words for TFIDF/Cosine in each of these queries were "reliable" and "diagrams" respectively. Many of the related but not entirely relevant search results for "reliable multicast" did discuss the topic of multicast distribution of data,

but did not address the issue of reliability, which in this space indicates that the work addresses the issues of congestion control and error correction. Similarly, for the query "software architecture diagrams", the TFIDF/Cosine system returned several documents that dealt with "software architecture", but none that address the construction and interpretation of software architecture diagrams as did the source document from which this query was drawn and four of the search results returned by Rosetta. In general, Rosetta demonstrates that even simple techniques based on the combined evidence of multiple references to documents provide excellent search performance in collections of scientific literature. Furthermore, the simplicity of these techniques provides additional support for the claim that these techniques may be easily mapped to other types of networked information.

CHAPTER 7

Related Work

7.1. Bibliometrics

For decades researchers have exploited link structures in collections of documents. Such research has been an ongoing topic in the field of bibliometrics. However, most of this work has focused on the structure rather than the content of collections of documents. In some of the earliest work Eugene Garfield [**27**] introduced the concept of citation indexing; a process in which the citations between documents were manually cataloged and maintained in much the same way as card catalogs of author, title, and subject so that beginning with any document a researcher might search through listings of citations or referrers, in essence traversing citation links either back through supporting literature or forward through the work of later researchers. Lawrence et al. later automated this process in CiteSeer [**28**], a Web-based information system that permits users to browse the citation links between documents as hyperlinks. In other work, Kessler [**33**] introduced the concept of bibliographic coupling for document clustering. In bibliographic coupling the degree to which two documents are similar is determined by the number of citations they have in common. For example, if we are looking at three documents, A, B, and C, if A and B cite ten of the same documents, A and C cite five

of the same, and B and C overlap with just two of the same bibliography entries, then the greatest similarity is likely between A and B, followed by A and C, and finally B and C. Turning this idea on its head, Small [**52**] exploits in-links rather than out-links. This process, called co-citation analysis determines the similarity of documents based on the number of citing documents two documents share. Finally, Salton in his customary ubiquity also weighs in this area. In [**48**] he argues that the set of index term extracted from documents may result in different search results for even only slightly different queries, because of the variation in word choice from one document to another and suggests that a more complete set of index terms might be found using the terms found in documents cited by a given document, documents citing a given document, and documents authored by the same author as the document in question.

7.2. Bibliometrics Applied to the Web

More recently, some, realizing the parallels between citations and hyperlinks have applied bibliographic coupling and co-citation analysis to the Web [**35**]. While citation indexing, bibliographic coupling, and co-citation analysis have been shows to be beneficial technologies, they are term-free operations; that is they are concerned with finding relationships without concern for the topics around which those relationships are based. As such they are not suitable for search applications, but rather for browsing or clustering. Salton's technique, while it is term based differs from mine in that it makes use of the entire text of neighboring documents.

Furthermore, his motivation appears to be better recall, where mine is primarily in improving precision. Finally, this technique was never to my knowledge demonstrated to have performance benefits in a retrieval setting. In fact, later work by Salton and Zhang [**51**] indicates that such a technique may degrade retrieval performance. Chakrabarti et al. [**15**] report a similar finding for classification tasks.

7.3. Web Link-Analysis Techniques

Advancing on the ideas in and around the field of bibliometrics, the HITS [**34**] and PageRank [**9**] algorithms were designed to rank Web pages on the basis of their popularity, where the popularity of a page is determined using a measure directly related to the number of other pages that link to that page. HITS is a post-processing ranking algorithm that identifies hubs and authorities in networks of Web pages seeded by the search results returned by a traditional search engine. A hub is a document that links to many other pages and an authority is a page to which many other pages link. The ideal hub is one that links to many authority pages. The ideal authority is one to which many hubs link. To find hubs and authorities, HITS takes the list of URLs returned in response to a query and gathers pages in the neighborhood of those URLs, by neighborhood I mean pages that are some constant number of links (in-links or out-links) away from the initial set of search results. It then uses the link information inherent in this subgraph of the Web to determine the hub and authority scores for each page

in the expanded set of search results. This approach has proved quite successful at identifying pages that are important in their service of one topic or another. Unfortunately, HITS is too time-consuming to realistically service search requests. In recognition of this problem, Brin and Page developed the PageRank algorithm [**9**] an algorithm that serves as the driving technology for the very popular Google search engine. PageRank works by pre-computing something similar to authority scores for all documents without regard for their topic. Since the PageRank values are pre-computed at indexing time, a simple table lookup for each page permits Google to rank the search results quickly. A recognized problem with this technique is that because the ranking of documents is heavily dependent on the topically impoverished PageRank measure, Google retrieves many documents that are important, but for reasons other than those identified in the query [**45**]. This is similar to the problem with ranking research papers merely on the basis of the number citations to them without a measure of relevance to queries. I do not mean to argue that Google does not provide satisfactory search results, rather I argue that it substitutes popularity for relevance, which works well for most queries, because by definition many queries target popular information. It is when a searcher is interested in that which is off the beaten path that Google is less helpful. The ideas behind PageRank are related to my work in that it pre-computes the importance of documents and uses that measure to rank search results. In addition, Google gathers anchor text used in reference to a document and incorporates the terms used therein with the body of terms used to index that document. However,

ranking is primarily driven by PageRank scores so search results are ordered not by the number of people that have described a document using the query words, but by the number of people who have described the document using any words at all. While I believe a document ranking approach more dependent on reference may help to solve some of the problems with the PageRank approach, a demonstration of this is outside the scope of this dissertation. Instead, as the initial stages of this body of work I focus on improving retrieval in scientific literature and in demonstrating that even the simplest of reference-based approaches can identify what documents are about better than the documents themselves. These findings not only contribute to information access in scientific literature, but by the generality of the technique, point to the possibility that greater topical precision can be achieve on the Web as well using such techniques, and as such make a small additional contribution in identifying a body of research that should be explored.

7.4. Use of Referential Text

Some researchers have touched on the idea of a reference-based approach to indexing and retrieval, but only to limited extent. However, what work has been done has provided positive results. Results that further substantiate my claim that further work in reference-based approaches to indexing the Web should be explored. McBryan with the World Wide Web Worm (WWWW) [**41**] 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 [**54**]. Very recently, other researchers have had even greater success with searches for homepages of people and organizations and other very broad queries using only anchor text, demonstrating superior retrieval performance to the content of documents [**4, 16**]. Though this work is limited to broad queries such as "Real Audio", "Best Buy", and "Disneyland", and "software", it does hint at the possibility of success with more specific searches for information such as those I have demonstrated with Rosetta. [**4**] is also limited in that the authors use only "expert" pages, that is those pages that are hub-like [**34**] in that that they link to several pages. An example of such a page is one listing several software packages for recording and playing MP3s. As a result, with this technique expert pages must first be identified, a process that my research indicates is probably not necessary given the ability of RDI to handle a great deal of noise in the input data. Both techniques are further limited in that they consider only anchor text and not surrounding text as well. As such will likely not be able to identify topics more specific than organizations or people, because anchor tends to contain the names of pages rather than an author's assessment of what that page is about as does the text surrounding the anchor text. See Figure 7.1 for an example.

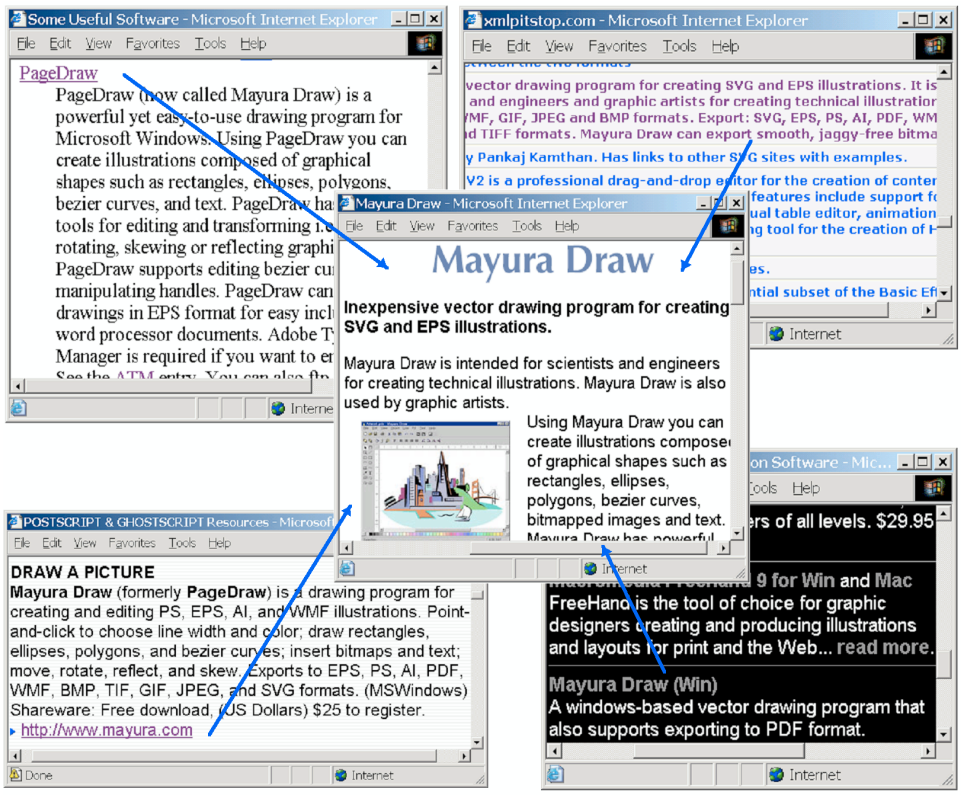


Figure 7.1. Four pages that reference www.mayura.com. The text used in the immediate vicinity of each link provides an excellent description of the target page, while the anchor text merely names the page.

7.5. Link Traversal

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 [**56, 44, 43**]. In related work, Davison [**19**] 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 [**20**] 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.

7.6. Use of Referential Text as Document Summaries

Other work views referential text from a very different perspective. Rather than using it as input to an information system, it uses such text as output, presented to users for the purpose of evaluating the relevance of search results. Amitay [**2**] has developed the InCommonSense system that collects anchor text and the text surrounding it from links embedded in clearly delimited paragraphs of text. InCommonSense uses machine learning techniques to identify the types of descriptions that people find most helpful in distinguishing between several results presented in response to queries and based on what it has learned selects one of several references to Web pages to display as the summary for document presented in a set of search results. The CiteSeer system permits similar interaction in allowing information seekers to view the contexts of citation to documents presented in search results. These contexts of citation are the type of referential text Rosetta uses in indexing documents.

7.7. Use of Referential Text in Classification

Researchers have also employed referential text in classification and categorization tasks. Notable among this work is that of Furnkranz [**25**] and of Chakrabarti et al. [**14**] two pieces of work I consider particularly relevant. Furnkranz uses anchor and surrounding text to classify documents from a university web page corpus based on the type of information it contains rather than its topic. The type of classifications he is interested in are similar to one type of meta-information I identified in Chapter4, that being the function of a document. Whereas I used overview, introductions, studies, etc. as examples, Furnkranz is interested in distinguishing student pages, faculty pages, project pages, etc. using a machine learning approach based on lexical features in referential text. Rosetta does not implement such a detector currently so I cannot compare my approach to his. More directly related to my work is that of Chakrabarti et al. In the work described in [**14**] the authors present an algorithm called ARC (Automatic Resource Compiler), which is an extension to Kleinberg's HITS algorithm and which they apply to the task of categorization of Web pages into broad Yahoo-like categories. The relationship between this work and my own is that Chakrabarti et al. use anchor text to enhance the topic specificity of HITS. The key difference between my work and this other than the obvious difference in domain of information is that, as mentioned above, HITS is not a feasible algorithm for search systems, because it is too slow. As a result, this technique is only applicable for offline categorization

of Web pages. My solution is a search solution and as such addresses a different class of problems. Furthermore, I have demonstrated the ability to identify what documents are about at a fairly specific topical level. This work, on the other hand, is intended for the type of broad categorizations inherent in ontologies such as Yahoo.

7.8. Most Closely Related Research

Finally, a few researchers have contrasted the benefits of anchor text with other text used in indexing. Cutler et al. [**18**] used a trial an error process to determine the optimal weighting of various types of text, including that found in title, header, strong, and anchor tags as well as body text. They found that in a small collection comparing retrieval over ten queries that anchor text and strong text should be weighted most heavily. In related work published at the same time I began publishing this work [**5**], Li describes an indexing approach for the Web [**39**] very similar to my own approach for scientific literature. The key difference between my work and that of Cutler et al. and Li is that it is far more extensive. I demonstrate a variety of indexing benefits in Chapter 4 including greater precision than content text, the ability to identify meta-information, and greater diversity of indexing terms, making it possible for more searchers to find what they need. Furthermore, in Chapter 3 I demonstrate Rosetta's ability to retrieve documents that are both on-point and significant over many queries and for topics of varying level of specificity. Neither Cutler et al. nor Li present such results. Cutler et al. do

not address the problem of significance of search results. Their algorithm does not factor in any measure that will promote significant documents. They are instead focused entirely on improving relevance. Li, while dealing with both relevance and significance, presents results that are hardly convincing. He does not test his system against other techniques running against the same data. Rather he tests it against search systems indexing much more and certainly thousands, perhaps millions of different documents, given what we now know about the differences in coverage of search engines [**37**]. Compounding the problem with Li's study is the fact that he uses a very small number of queries (10) and many of the queries reflect those similar to the homepage finding work discussed above, notable examples are "Yahoo" and "Microsoft".

Overall, while several researchers have demonstrated some benefit in using anchor text for a variety of applications no one has provided an extensive treatment of the topic, demonstrating the ability to locate not simply well-named types of information such as company home pages, but also information on specific and less well-named topics. Furthermore, my research is the first to demonstrate the power of a reference-based technique in identifying information that is both relevant and significant. In addition I have demonstrated the robustness of the technique in its ability to handle large volumes of noise in the input data. Finally, my work broadens our understanding of referential text of which anchor text is only one type and indicates that a reference-based approach in many mediums may provide superior performance over content-based approaches in text retrieval systems.

CHAPTER 8

Future Work

Like many projects a dissertation is as much a beginning as it is an end to a body of work. As a result it opens just as many questions as it provides answers. It has been my intention to think big about the concept of indexing by reference, and rather than focus on some small aspect of this work to provide experimental evidence that substantiates the value of this approach to indexing and identify a range of future paths of exploration that will refine the finding I have presented here. In particular, the work presented here demonstrates that a comparison of multiple references to documents precisely identifies what makes documents valuable as sources of information. Furthermore there is some evidence to suggest that such an analysis also identifies the way people naturally write and talk about the ideas in a body of information. I believe expanding the work to other data in a referential structure of some kind exists will yield better indexing and retrieval in a variety of information mediums outside of scientific literature. In addition, I believe the collaborative query interface to Rosetta is just the beginning of new paradigm of access to information.

While the focus of this dissertation has been on improving access to scientific literature, the indexing and retrieval techniques presented herein map quite easily

to the Web where people seem to be compelled to demonstrate their knowledge of the topic about which they are speaking. This occurs for a variety of reasons including a genuine desire to point people to information they may find useful. Others cite other Web documents because they wish to demonstrate their own level of knowledge in a particular subject area. Still others do so out of obligation to a class of students, a group of employees, employers, clients, or colleagues, or some other group of people with whom they regularly interact. Furthermore, given the integration of the Web with nearly everything we do, from ordering prescriptions at Walgreens.com to chatting with friends, it is not difficult to imagine a world in which nearly every document written is created on-line and is directly linked to many others that are related in some way. And if the Web is any indication of what is to come, most useful documents will be linked to from a number of others in which the authors of referenced them using simple accurate labels for the information they contain. After all, even the simplest of documents, if I may use the term so loosely, the instant message often contains hyperlinks cut and pasted from a browser with a brief note describing them. The technology I have described in this dissertation is already broadly applicable to many Web-related information sources. As more and more information is generated on-line and old paper files are transfer to electronic forms, a technology that exploits the descriptions for documents people create in the natural course of their activities will become increasingly valuable. More importantly, with the growing volume of information available on-line, we need a technology that is able to decipher the tasks for which

documents are useful and the degree of that utility. The reference-based technology I have herein described that compares and contrasts references to documents acquires with great precision the labels people use for the information they contain. Because this technology is by design general-purpose, requiring only several bits of text written in reference to each document, it is directly applicable to Web pages in which people are free in their criticism and praise of the writing of other Web authors.

A Web search engine is an obvious future application of this technology, but I believe its applicability is broader than simply building search engines. This technology has great potential for tracking and making available to others important data concerning the ways in which people use information. After all, as others in addition to myself have noted [**1, 57, 8**], the way that people have used a piece of information in the past makes the best basis on which to determine how that information will be useful to people in the future. I believe an RDI approach to indexing is particularly well suited to an area particularly lacking in effective information access solutions, that being the area of knowledge management in large organizations. Since much of the recent work in IR has focused on the publicly available Web, less attention has been paid to the knowledge management needs of companies and organizations. Large organizations with many employees maintain large amounts of information including research reports, performance reports, market analyses, competitor analyses, and many other types of information that

represent what the company as a whole knows. Unfortunately much of this information is poorly organized and for all practical purposes inaccessible to most members of the organization. In a large part this is due to the fact that effective information organization is a time-consuming and difficult task when done by hand and one that results often imprecise indexing and retrieval when done automatically. As one further application of the technology I have described in this dissertation, I plan to implement a system that tracks intra-organization email and other forms of communication such as chat applications, and indexes attachments, the targets of URLs, and other documents in some way passed between members of the organization with short textual descriptions indicating how they are useful. Such interactions occur for a variety of reasons in an organization, some of which are actually for the purpose of accomplishing work. Aside from the challenge of distinguishing work-flow from pointers to humorous Web-sites and the like are the challenges such as matching multiple descriptions to the same report, for example, as it is passed from one employee to another. However, I do not believe these challenges are insurmountable and when completed such a system should prove an invaluable resource that will allow organizations to not only track the flow of information through their organizations in new and useful ways, but prevent the type of redundant and costly work on which organizations waste a great deal of money, because their employees are not aware of the efforts of their co-workers.

Other future applications of this technology arise from applying the lessons learned in developing the collaborative query interface component of Rosetta. As a

long-term goal I envision a system that effectively interacts with people in research processes to find, suggest, summarize and organize the information they read and create to improve learning and the speed with people are able to acquire knowledge. This body of work will integrate my dissertation work with work on just-in-time information access [**11, 10, 12, 17**] to create technology that streamlines research-oriented information gathering tasks for a range of people including high school students working on term papers to executives researching the competition for their company's products. For example, imagine a student writing a paper on the politics of crisis and a collaborative system embedded in the Microsoft Office suite of tools. Imagine that the student plans to use as a case study, the key players in the Kennedy Administration during the Cuban Missile Crisis. To begin her inquiry she submits the query "cuban missile crisis" from an interface within Explorer. In the list of documents retrieved she finds one that summarizes the members of the Kennedy Administration. Believing this document a good place to start the student requests a summary of the document. The system takes her to a page listing the important features of that document based on descriptions extracted from pages that link to that document. Among the features listed are the names of administration officials. The student is familiar with the roles of Bobby Kennedy, Lyndon Johnson and several others, but less so with that of Dean Rusk, Secretary of State at the time. Therefore, as her next step she chooses to follow a link from "Dean Rusk" to a query for information about him, a query for which the system filters out documents such as the home page for the Dean

Rusk library or something that do not concern Dean Rusk and his involvement with the Cuban Missile Crisis. Reading over the resulting documents she begins to take notes in Word. One document alludes to the idea that though his role was originally viewed as insignificant, Mr. Rusk's work behind the scenes was critical to the successful resolution of the crisis; she jots down a note to this effect. After taking several notes on Mr. Rusk, she is ready to dig a little deeper into some of the more interesting ideas she has come across. Anticipating the possibility of this step, the system has provided links to more information on each of the notes she has written. The student chooses the link to more information on Mr. Rusk's work behind the scenes, she views a list of documents retrieved through a query automatically constructed and submitted based on her note. To make it easier for the researcher to process the search results the system automatically categorizes them based on the most frequently occurring common features they exhibit. The features will likely define sub-categories according to key interactions between Mr. Rusk and other cabinet members, and will be organized into well-named categories using technology similar to the collaborative query interface in Rosetta. Following this step the student continues to use the tools provided in this interface, moving among the tasks of gathering, organizing, and writing in a variety of ways until her paper is complete.

The research agenda I envision for the future represents an integration of various techniques into a system that blurs the line between information gathering and creation to such an extent that users will move seamlessly through an information

space in an efficient and effective manner. Through this work I will endeavor to combine the indexing and retrieval power of RDI with the summarizing capabilities of referential text [**2**] in technology embedded in everyday applications [**10, 23, 17**].

References

[1] Brian Amento, Loren G. Terveen, and William C. Hill. Does "authority" mean quality? predicting expert ratings of Web documents. *Research and Development in Information Retrieval*, pages 296–303, 2000.

[2] Einat Amitay and Cecile Paris. Automatically summarising web sites - is there a way around it? In *Proceedings of the ACM Ninth International Conference on Information and Knowledge Management*, pages 173–179, 2000.

[3] Nicholas J. Belkin, C. Cool, D. Kelley, S.-J. Lin, S. Y. Park, J. Perez-Carballo, and C. Sikora. Interative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing and Management*, 37:403–434, 2001.

[4] 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.

[5] Shannon Bradshaw. Reference directed indexing: Attention to the description people use for information. Master's thesis, The University of Chicago, Chicago, IL USA, December 1998.

[6] Shannon Bradshaw and Kristian Hammond. Constructing indices from citations in collections of research papers. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, Washington, D.C., November 1999.

[7] Shannon Bradshaw and Kristian Hammond. Automatically indexing documents: Content vs. reference. In *Proceedings of the Sixth International Conference on Intelligent User Interfaces*, San Francisco, CA, January 14–17 2002.

[8] Shannon Bradshaw, Andrei Scheinkman, and Kristian Hammond. Guiding people to information: Providing an interface to a digital library using reference as a basis for indexing. In *Proceedings of the Fourth International Conference on Intelligent User Interfaces*, New Orleans, LA, January 9–12 2000.

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

[10] J. Budzik, S. Bradshaw, X. Fu, and K. J. Hammond. Clustering for opportunistic communication. In *Proceedings of the Eleventh International World Wide Web Conference*. ACM Press, 2002.

[11] Jay Budzik, Shannon Bradshaw, Xiaobin Fu, and K. Hammond. Supporting online resource discovery in the context of ongoing tasks with proactive assistants. *International Journal of Human-Computer Studies*, 56(1):47–74, January 2002.

[12] Jay Budzik and Kristian J. Hammond. User interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 2000 International Conference on Intelligent User Interfaces*, New Orleans, LA, 2000. ACM Press.

[13] James P. Callan, W. Bruce Croft, and John Broglio. TREC and tipster experiments with inquery. *Information Processing and Management*, 31(3):327–343, 1995.

[14] 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.

[15] Soumen Chakrabarti, Byron E. Dom, and Piotr Indyk. Enhanced hypertext categorization using hyperlinks. In Laura M. Haas and Ashutosh Tiwary, editors, *Proceedings of SIGMOD-98, ACM International Conference on Management of Data*, pages 307–318, Seattle, WA USA, 1998. ACM Press, New York.

[16] 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.

[17] Andrew Crossen, Jay Budzik, Mason Warner, Lawrence Birnbaum, and Kristian J. Hammond. XLibris: An automated library research assistant. In *Proceedings of the Fifth International Conference on Intelligent User Interfaces*, pages 49–52, Santa Fe, NM, 2001.

[18] Michael Cutler, Yungming Shih, and Weiyi Meng. Using the structure of html documents to improve retrieval. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems (NSITS'97*, pages 241–251, December 1997.

[19] 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.

[20] 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.

[21] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[22] Adam Ferrari. JPVM: Network parallel computing in Java. *Concurrency: Practice and Experience*, 10(11-13):985–992, 1998.

[23] David Franklin, Shannon Bradshaw, and Kristian J. Hammond. Jabberwocky: You don't have to be a rocket scientist to change slides for a hydrogen combustion lecture. In *Proceedings of the Fourth International Conference on Intelligent User Interfaces*, pages 98–105, New Orleans, LA, January 9-12 2000.

[24] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[25] Johannes Furnkranz. Exploiting structural information for text classification on the WWW. In *Intelligent Data Analysis*, pages 487–498, 1999.

[26] E. Garfield. *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. The ISI Press, Philadelphia, PA, 1983.

[27] Eugene Garfield. Citation indexes for science. *Science*, 122:109–111, July 1955.

[28] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, 1998.

[29] Karen Sparck Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information retrieval: Development and comparative experiments - part 2. *Information Processing and Management*, 36(6):809–840, 2000.

[30] Steve Jones, Sally Jo Cunningham, and Rodger J. McNab. Usage analysis of a digital library. In *ACM DL*, pages 293–294, 1998.

[31] 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.

[32] M. M. Kessler. Technical information flow patterns. In *Proceedings of the Western Joint Computing Conference*, pages 247–257, Los Angeles, CA, 1961.

[33] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[34] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[35] R. R. Larson. Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace. In *Proceedings of the Annual American Society for Information Science Meeting*, pages 71–78, Baltimore, MD, 1996.

[36] 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.

[37] Steve Lawrence and C. Lee Giles. Searching the World Wide Web. *Science*, 280:98–100, 1998.

[38] Anton Leuski and James Allan. Improving interactive retrieval by combining ranked lists and clustering. In *Proceedings of RIAO 2000*, pages 665–681, April 2000.

[39] Yanhong Li. Toward a qualitative search engine. *IEEE Internet Computing*, 2(4):24–29, July/August 1998.

[40] H. P. Luhn. The automatic derivation of information retrieval encodements from machine-readable texts. *Information Retrieval and Machine Translation*, 3(2):1021–1028, 1961.

[41] 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.

[42] http://www.medline.com. A database of journal articles published by the U.S. National Library of Medicine.

[43] 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.

[44] 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.

[45] 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.

[46] J. J. Rochio. *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter Relevance Feedback in Information Retrieval. Prentice Hall, Englewood Cliffs, N. J., 1971.

[47] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[48] Gerard Salton. Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, 10(4):440–457, October 1963.

[49] Gerard Salton and Christopher Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[50] 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.

[51] Gerard Salton and Y. Zhang. Enhancement of text representations using related document titles. *Information Processing and Management*, 22(5):385–394, 1986.

[52] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265–269, 1973.

[53] Karen Sparck-Jones. A statistical interpretation of term specificity and its application to retrieval. *Journal of Documentation*, 28(1):11–20, March 1972.

[54] Ellen Spertus. *ParaSite: Mining the Structural Information on the World-Wide Web*. PhD thesis, MIT, February 1998.

[55] 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.

[56] 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.

[57] Loren Terveen, William Hill, Brian Amento, David McDonald, and Josh Creter. PHOAKS: A system for sharing recommendations. *Communications of the ACM*, 40(3):59–62, 1997.

[58] J. H. Westbrook. Identifying significant research. *Science*, 132, October 1960.