NORTHWESTERN
UNIVERSITY

Computer Science Department

Technical Report
NWU-CS-01-7
2001

Document Indexing Vocabularies: Reference vs. Content

Shannon Bradshaw

Abstract

A search engine is only as good as its ability to pair people with the information they need. Current search technology is moderately successful when asked to satisfy simple information needs, but inadequate for complex, research-oriented tasks in which fine distinctions between documents are necessary. As a means of enhancing the ability of researchers to find the information they need I present a new indexing technique based on the words used in reference to documents. While others have also suggested the use of referential text (hypertext) in indexing documents, my approach is novel in that it identifies not just the most recommended documents, but also the subjects for which these documents are highly recommended. It is therefore better able to distinguish popular but irrelevant documents from those which will be more useful to an information seeker. In this chapter I present a study of the indexing vocabulary provided by reference in a collection of academic research papers. Using content as a baseline for comparison, I measure reference against the metrics of topical precision, identification of meta-information, and term diversity. The results of this study 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.

Document Indexing Vocabularies: Reference vs. Content

Shannon Bradshaw
Department of Computer Science
Northwestern University
1890 Maple Ave
Evanston, IL 60201

TR-NWU-CS-01-07

ABSTRACT

A search engine is only as good as its ability to pair people with the information they need. Current search technology is moderately successful when asked to satisfy simple information needs, but inadequate for complex, research-oriented tasks in which fine distinctions between documents are necessary. As a means of enhancing the ability of researchers to find the information they need I present a new indexing technique based on the words used in reference to documents. While others have also suggested the use of referential text (hypertext) in indexing documents, my approach is novel in that it identifies not just the most recommended documents, but also the subjects for which these documents are highly recommended. It is therefore, better able to distinguish popular but irrelevant documents from those which will be more useful to an information seeker. In this chapter I present a study of the indexing vocabulary provided by reference in a collection of academic research papers. Using content as a baseline for comparison, I measure reference against the metrics of topical precision, identification of meta-information, and term diversity. The results of this study 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.

1. Introduction

A search engine is only as good as its ability to pair people with the information they need. Though there remains much work to be done, the average user of existing web search engines reports a moderate degree of success in finding the information he needs. In considering the success of such technology, it is important to note that the majority of queries for which search engines perform well are requests for information in which many people are interested. Indeed, the most successful search engines are based on indexing and retrieval techniques that prefer popular web pages as determined by the number of paths composed of one or more hyperlinks leading to that page from others [6]. For many people the fact that they are interested in the same celebrity, computer game, or digital camera as thousands of other people means that from a simple two or three word query a search engine is often able to correctly guess the information desired. In addition, since many of these queries are satisfied by any one of as many as thousands of pages the problem of locating needed information is greatly simplified.

In contrast, existing search technology performs poorly for people performing research-oriented tasks. Whether a middle school student working on a science project or a management consultant trying to improve the profitability of his client, the information needed by researchers is usually somewhat obscure. In addition, where thousands of documents may satisfy the casual information seeker, it is likely that only a handful will provide the researcher with what he needs. The reason for this is that researchers are often interested in a particular viewpoint, set of results, or some other finely distinguished piece of information within their subject of interest. These conditions place far more stringent requirements on an information retrieval system. Especially since the search behavior of researchers does not appear to differ significantly from that of the average user of web search engines [8]. In order to pair researchers with the information they need, a search engine must be able to distinguish the relative value of documents that might be considered equally valuable using a broad measure such as raw popularity.

For any query, many documents are relevant in that they address the topic identified in the search to some extent. However, far fewer contribute the particular information in which a researcher may be interested. Therefore, an information system should index documents using identifiers for the contributions they make to a body of knowledge, excluding identifiers for less useful information they may contain. Indexing vocabulary should identify not only the subject areas to which a document contributes but also any meta-information that may further distinguish the utility of that document for a given information seeker. For example, an information system that is able to distinguish introductory material on a particular subject from that which is more advanced will be better able to serve researchers, because it can filter search results according to their level of knowledge. Finally, human language is rich and expressive and researchers with varying levels of knowledge will tend to use different words to describe the information in which they are interested. Therefore, the vocabulary with which documents are indexed should be diverse, reflecting many ways of describing the important features of each document [4].

Descriptions that provided this type of information about documents are inherent in the structure of most on-line information. Reference, whether in the form of hypertext or traditional citations, pinpoints the contribution of a document with just a sentence or two. For example with the sentence, "…XML QL [9] is a proposed query language for XML with many of the features needed for a shared, distributed data

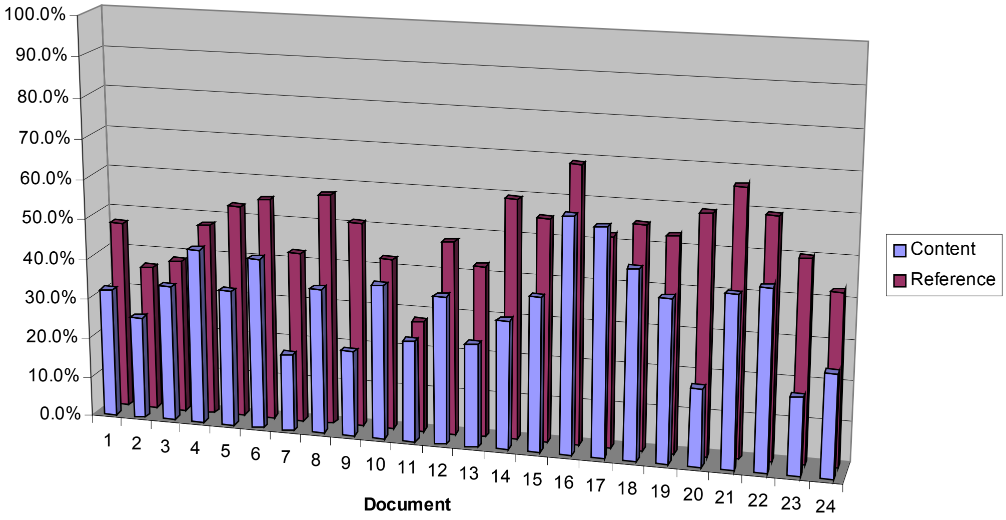

Figure 1: A pair-wise comparison of reference to content against a metric defined by the percentage of index words drawn from each source that identify the key subjects of a document.

store."[1] Karavanic describes a document by Deutsch et al. The language used in such sentences precisely identifies the subjects on which cited documents contain useful information. In addition, referrers typically identify other distinguishing features of a document such as whether it is an introduction, overview, etc. Finally, multiple references to a document provide the diversity of language inherent in the perspectives of multiple authors. Therefore, I believe reference provides an excellent source of identifiers for indexing documents.

In this chapter, I present a study of reference as a basis for indexing a collection of scientific literature. The results of this study demonstrate that the words of those who cite a document provide the fine distinctions necessary to recognize subtle differences between related documents and do so using language that is rich enough to match the queries of many different people for the same information. This study indicates that reference may provide the indexing information necessary to support a more sophisticated search technology than is possible with current approaches to indexing – a technology that will better serve the more complex information needs of researchers.

2. The Study

To perform this study I collected approximately 30,000 of the documents maintained by ResearchIndex [10] and indexed them by both content and reference. Since most search engines rely heavily on the words used within a document for indexing and retrieval, I use indices extracted from content as a baseline for determining the quality of indices drawn from reference. I compared these two sources of indexing vocabulary against the three metrics of subject precision, identification of meta-information, and language diversity. For references I used windows of text surrounding citations that are approximately 50 words in length. I used the same indexing technique for both content and reference and based this technique on traditional IR methods so that I was looking at content as it is typically used. I weighted the indices for each document using a standard TFIDF metric [12].

From the collection, I gathered a sample of 24 documents to study. These documents were required to meet 2 restrictions, but were otherwise selected at random. First, I required that each document had been cited at least 20 times to ensure that there was enough text with which to index a document in the reference database. Second, I required each document to contain a list of keywords specified by the author. I imposed this restriction so that I did not introduce bias toward reference by determining the important features of documents myself.

With the sample set of documents chosen, I identified the key features of each document that distinguish it in a way information seekers would find useful. Using the keywords listed by the authors to guide my decisions, I determined the importance of a paper using the abstract, introduction, and other content of the document. I identified both the subjects to which documents contribute as well as distinguishing meta-information. For subjects I was not looking for broad research area identifiers, but was instead looking for specific contributions within a particular area. For example, one paper I considered falls within the area of "artificial life"; however, the key contribution of that paper is work on "evolution of cooperative communication". I chose the latter to the exclusion

[1] From Karavanic, K. *Experiment Management Support for Parallel Performance Tuning*. Doctoral Dissertation. University of Wisconsin. 1999. in reference to Deutsch A., M. Fernandez, D. Florescu, A. Levy, and D. Suciu. *A query language for XML*. Technical report, AT&T Labs, 1998.

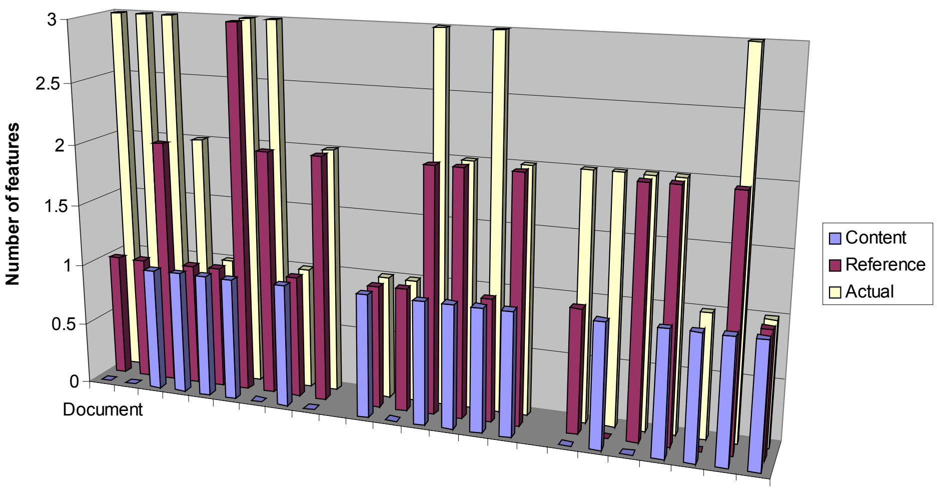

Figure 1: A comparison of content and reference measuring the number of known meta-information features identified for each document.

of the former as a key feature of that document. For meta-information I looked for non-topical features of documents that serve to fine-tune their value to an information seeker. For example, important meta-information might include the type of contribution a document makes whether that is an algorithm, proof, or software package.

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. I considered only those indices that will position a document in a set of query results so that an information seeker will be likely to view it. Most people, including researchers [7], are unwilling to exert much effort when searching for information. They submit queries no longer than two words [7, 16], do not look past the first page of results returned in response to a query [16, 17], and are unwilling to submit more than one query for the same information [13]. Given this behavior only the most heavily weighted indices will cause an information seeker to actually see a document. After sampling the distribution of weights for indices extracted from both content and reference I determined that by evaluating the 50 most heavily weighted terms from each source for each document, I would be assured of considering only terms that are likely to place that document within the first page of query results.

3. Study Results

I found that the indices extracted from reference do provide an excellent source of identifiers for the important information a document contains. Better identifiers, in fact, than those drawn from content. Reference significantly outperformed content in the precision with which it identified key subjects, in the amount of important meta-information it identified, and in the diversity of indices it supplied that name these features.

3.1 Identifying Subjects

In measuring the precision with which each indexing vocabulary identified the key subjects of documents, I considered an average of 4.4 subjects per document. The most I considered for any one document was 7 and the least was 2. Some of the key subjects I identified 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. In each set of words I looked for those that identified key subjects for the documents they indexed. I marked as subject identifiers words used within the text from which the indices were drawn to name a feature either as part of a group (i.e. “quality” in “quality of service”) or singularly (i.e. “QoS”). Words marked as subject identifiers were those from any part of speech that were used within the text from which it was drawn to name one of the key subjects. For example for the topic “contact constraints” the words “contact”, “touch”, and “touching” were all considered valid identifiers. 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 key subject for a document, while 50.5% of indices from reference identified the same subjects. Comparing the number of subject identifiers on a document-by-document basis, I found that the mean paired difference was 15.6% 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 indices originating in content and reference, about 8 more of those coming from reference identified a key subject for each document I evaluated. Figure 1 shows the relative precision of the indices drawn from content and reference and a percentage of the total number indices considered for each document.

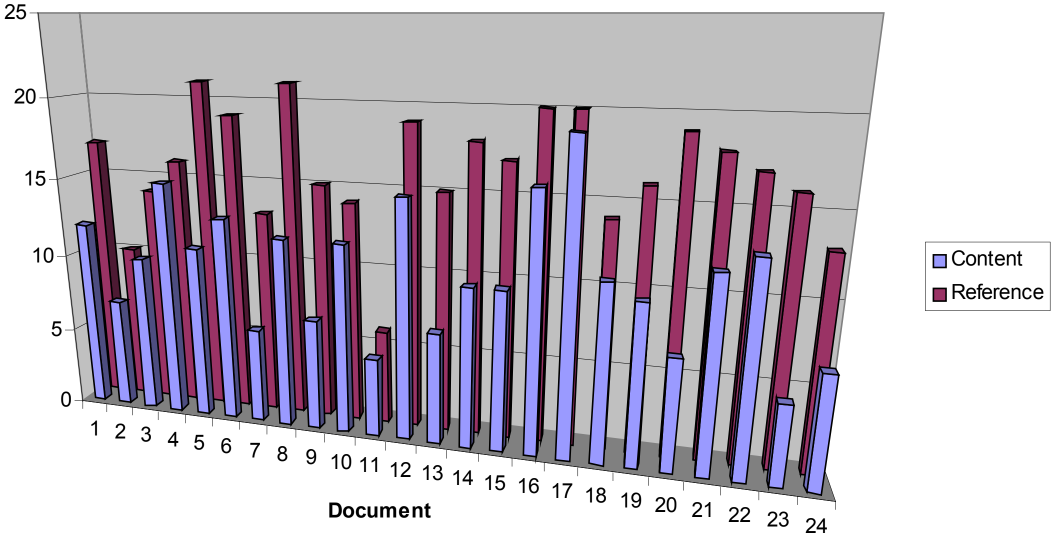

Figure 2: A comparison of content and reference on the basis of the number of unique ways of describing each document.

3.2 Identifying Meta-Information

Reference also performed well in identifying important meta-information concerning documents. In this phase of the study, I was interested in indices that identified useful extra-topical features such as the type of contribution made by a document. For example, one document contained important study results; another contributed a new algorithm in the area of mobile computing. I identified an average of 2 pieces of meta-information per document, but there were 2 documents for which I could find no useful meta-information. As with subjects, I marked as meta-information identifiers words that named such as feature either singularly (i.e. "algorithm") or as part of a group of words (i.e. "study" for "study results"). I found that content indices identified the meta-information for a document in only 23% of the cases, while reference indices identified all meta-information for 50% of the documents I considered. Comparing the relative performance per document, reference 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 content identified more meta-information. Figure 2 shows the relative performance of content and reference in identifying useful meta-information on a document-by-document basis.

3.3 Measuring Indexing Language Diversity

Finally, to test the degree to which reference meets the indexing needs imposed by a diverse search vocabulary, I compared the diversity of indices drawn from reference to those from content. In this evaluation, I looked at the number of different ways in which a document as a whole was described. I categorized the indices for each document based on the way they were used in the text from which they were drawn. 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 terms used to identify key subjects I found that the average number of unique identifiers per document originating in content was 10.5 while from reference I 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 3 charts the difference between content and reference as sources of unique identifiers for key subjects.

The greater diversity of indices arising from reference is the result of many authors citing the document in the context of their own work. 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 words drawn from reference create a larger target for searchers to hit than those drawn from content. In other words, the context in which a searcher is working has a much better chance of matching the context of one or more citers of a document than it does of matching only the context of its author.

3.4 Another Look at Subject Identifiers

Having reported on the three metrics of primary interest in this study, I return to the issue of precision. Combining the indices for both key subjects and meta-information 52.8% of those from reference identified a key document feature compared to 35.8% of those from reference. The question then remains -- what did the other indices identify? For both content and reference the overwhelming majority of these indices (approximately 70% for both sources) identified details about a paper in addition to the key subjects and meta-information I had determined before beginning the 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 indices 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 indices I looked at the source from which the words were drawn. My goal was to determine why they appeared in the list of indices for a document. I found that over 90% of the words drawn from reference were weighted heavily because several researchers used that word to identify a document feature they found important. For content 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 indices from content serve as good indices for a document I compared them to those drawn from reference for a sample of 10 documents used in this study. I found that the indices drawn from content identify only 63% the features referrers considered important. While this result does not necessarily mean that the remaining indices from content poorly identified what is useful about the documents I evaluated, it does mean that on average over one-third of the content indices for each document identified concepts that not one of at least twenty authors identified as important.

4. 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.

The notion of using links to a document as recommendations of that document is not new. Several researchers have addressed this issue in Web documents [9, 2, 11, 10, 15] and researchers in biblio-metrics have worked with link structure for decades [5, 14]. Indeed, others have even suggested the use of referential text such as hypertext as a means of determining the relevance of documents to queries [2]. However, such approaches do so almost as an afterthought. PageRank [2] in particular determines the value of a document without consideration of its subject. Search engines based on PageRank retrieve the most popular documents even if they are only loosely associated with the words in the query either through hypertext or through the content of the documents themselves. What I present as a contribution is the use of reference as a means of determining not just the most recommended documents, but also the subjects for which these documents are highly recommended. I believe this will provide improved retrieval through matching descriptions of information needed (queries) with the documents that have been most often recommended as providing that information.

In performing the study I present here it was my goal to understand how well reference might serve in support of the information needs faced by researchers. I have demonstrated that reference precisely identifies the useful information documents contain 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 demonstrate that reference better captures the essence of a document and motivates the development of an information retrieval system in which documents are indexed by reference. In the chapters that follow, we describe such a system and demonstrate that it does indeed perform better than a similar system in which documents are indexed by content.

5. REFERENCES

[1] Bradshaw, S., A. Scheinkman, and K. Hammond. Guiding People to Information: Providing an Interface to a Digital Library Using Reference as a Basis for Indexing. In *Proceedings of IUI 2000,* New Orleans, LA, Jan 9-12, 2000.

[2] Brin, S. and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of WWW '98*. Brisbane Australia, April 1998.

[3] Chakrabarti, S., B.E. Dom, D. Gibson, J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagoplan, and A. Tomkins. Mining the Link Structure of the World Wide Web. IEEE Computer, 32(8), 60-67, 1999.

[4] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, 30(11), 964-971, 1987.

[5] Garfield, E. *Citation indexing: its theory and application in science, technology and humanities*. The ISI Press, Philapdelphia, PA, 1983.

[6] The Google Search Engine. http://www.google.com.

[7] Jones, S., Cunningham, S. J., and McNab, R. An Analysis of Usage of a Digital Library. *Proceedings of ECDL '98*. Heraklion Crete Greece, Sept 1998.

[8] Jones, S., Cunningham, S. J., McNab, R., and Boddie, S. J. A Transaction Log Analysis of a Digital Library. International Journal on Digital Libraries, 3(2):152-169, 2000.

[9] Kleinberg, J. Authoritative sources in a hyperlinked environment. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 668-677, January 1998.

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

[11] Pirolli, P., J. Pitkow, R. Rao. Silk from a sow's ear: Extracting usable structures from the Web. In: Proceedings of the ACM-SIGCHI Conference on Human Factors in Computing Systems, Vancouver, British Columbia, Canada, 1996.

[12] Salton, G. and Buckley, C. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5), 513-523, 1988.

[13] Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33(3), 1999.

[14] Small, H. 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.

[15] Spertus, E. ParaSite: Mining structural information on the Web. In *Proceedings of the 6th International World Wide Web Conference*, Santa Clara, CA, 1997.

[16] Spink, A. A user centered approach to the evaluation of Web search engines: An exploratory study. *Information Processing and Management*. 2001.

[17] Spink, A., D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 53(2): 226-234,2001.