Automatically Indexing Documents: Content vs. Reference

Shannon Bradshaw and Kristian Hammond
Department of Computer Science
Northwestern University
Evanston, IL 60201 USA
{bradshaw, hammond}@cs.northwestern.edu

Abstract
Authors cite other work in many types of documents. Notable among these are research papers and web pages. Recently, several researchers have proposed using the text surrounding citations (references) as a means of automatically indexing documents for search engines, claiming that this technique is superior to indexing documents based on their content [1,2]. While we ourselves have made this claim, we acknowledge that little empirical data has been presented to support it. Therefore, in the limited space available we present a terse overview of a study comparing reference to content as bases for indexing documents. This study indicates that reference identifies the value of documents more accurately and with a greater diversity of language than content.

Keywords
Indexing precision, term diversity, reference-based indexing.

INTRODUCTION
A search engine is only as good as its ability to pair people with the information they need. More specifically the quality of such a system is best measured by the success with which it pairs queries with useful documents. For any query, many documents are relevant in that they address the topic identified in the search to some extent. However, far fewer contribute important information to a body of knowledge. Documents making important contributions are far more useful to people than those that are merely relevant. Therefore, an information system should index documents using identifiers for the contributions they make to the exclusion of other information they may contain.

Precise indexing is the foundation of a good information retrieval system. But no matter how precisely a system identifies the importance of documents, if the identifiers used to index them do not match queries for the information they contain then the system will perform just as poorly as one in which documents are indexed imprecisely. Because human language is so rich and expressive, people use many different words to describe the same concepts in queries to search engines [3]. Therefore, the vocabulary with which documents are indexed should be diverse, reflecting many ways of describing the important features of each document.

With the growth of Internet access during the past decade, much of the information people use to research one idea or another is available on-line. An important part of this structure and indeed part of the Internet as a whole is the mechanism of reference, whether in the form of hyperlinks or traditional citations. In citing another document an author typically pinpoints the value of that document with just a sentence or two in which the author describes the work in relation to his own. The language used in these sentences is not only precise, but multiple citations to a document provide the diversity of language inherent in the perspectives of multiple authors. Reference, therefore, provides an excellent source of identifiers for indexing documents.

Since most search engines rely heavily on the words used within a document for indexing and retrieval, an important evaluation of reference is how it fares when compared to content as a means of identifying the important features of documents. In this paper, we argue that reference is superior to content as a basis for indexing. As evidence we present a comparison of the indexing vocabularies extracted from content and reference in a collection of Computer Science research papers.

IUI'02, January 13-16, 2002, San Francisco, California, USA.

The Study
For this study we used the metrics of topical precision, coverage for meta-information, and term diversity. By topical precision we mean the success with which the indices for a document identify the subjects in which a paper makes contributions, to the exclusion of identifiers for other subjects a document may address. Meta-information coverage is the degree to which the indices for a document identify extra-topical features such as the type of contribution a document makes (i.e. algorithm, study results). Term diversity measures the number of different identifiers for the same idea within the indices for a document. The first two metrics measure the success with which an indexing vocabulary identifies the value of a document. The third measures the size of the target information seekers must hit in searching for that information. These properties are fundamental to the success or failure of an information system.

We collected approximately 30,000 of the documents maintained by ResearchIndex [5] and indexed them by both content and reference. For references we used windows of text surrounding citations that are approximately 50 words in length. We used the same indexing technique for both content and reference and based this technique on traditional IR methods so that we were looking at content as it is typically used. We weighted the indices for each document using a standard *tfidf* metric in which important indices are those that occur frequently in the text used to index a document and rarely in the text used to index other documents.

From the collection, we gathered a sample of 24 documents to study. These documents were required to meet 2 restrictions, but were otherwise selected at random. First, we 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. This is actually a small number of citations for a document people will find useful and would likely be reached during the first year or two following publication. Second, we also required each document to contain a list of keywords specified by the author. We imposed this restriction so that we did not determine the important features of documents solely based on our own opinions, and thereby increase the risk of introducing bias toward reference.

With the sample set of documents chosen, we identified the ways in which each document contributes to the field of Computer Science. Using the keywords listed by the authors to guide our decisions, we determined the importance of a paper using the abstract, introduction, and other content of the document. In a further effort to avoid introducing bias toward reference, we used only the content of documents to determine key contributions. We identified both subject areas in which a document makes contributions and meta-information.

We then evaluated the degree to which each indexing vocabulary identified these features. Given the typical search behavior of people [4,7,8], only the most heavily weighted indices will cause an information seeker to actually see a document. After sampling the distribution of term weights in both the content and reference databases we determined that by evaluating the 50 most heavily weighted terms for each document, we would be assured of considering only terms that are likely to place it within the first page of query results.

With all the sample data collected we evaluated the quality of the indexing vocabularies from content and reference. In each set of indices we looked for words that identify important and distinguishing document features. We marked as feature identifiers, terms that name a contribution either as part of a group of words (i.e. "quality" in "quality of service") or singularly (i.e. "QoS"). To verify feature identifiers we required that usage be demonstrated in either the document itself or in reference to that document, regardless of whether the word originated in content or reference.

STUDY RESULTS

Authors cite documents using language that describes other work in relation to their own; in the process, they summarize the importance of that work. Documents are far more complex, often requiring much text that does not directly identify the importance of a document. Our analysis of the precision with which content and reference identify the important topics of documents supports this reasoning. In evaluating each indexing vocabulary we found that on average 35 % of content indices identify at least one key feature, compared to 51% of indices from reference. The mean difference was 16% with a standard deviation of 10% and a 90% confidence interval of $\pm$ 3.5%. Further analysis indicates that most of the misleading indices drawn from reference identify related work, and result from multiple documents being cited near one another. Using a simple filter we have been able to eliminate much of this text in the general case. No such solution is evident for the misleading indices drawn from content.

Coverage for Meta-information

In addition to important subjects, authors also identify distinguishing extra-topical features using phrases such as "good overview". This type of information is difficult to extract from the content of documents. In fact, it is usually only through the collective opinion of the research community that a document becomes known as a "good overview", "good introduction", etc. Testing this hypothesis, we found that content indices identified all the meta-information for a document only 23% of the time, while reference indices identified all meta-information for 50% of the documents we considered. Comparing relative performance per document, reference identified more meta-information for 64% of the documents and identified the same number of features for 27% of the documents, leaving only 2 documents for which content identified more meta-information.

Vocabulary Diversity

As evidenced so far, the individual perspectives of many citing authors work together to provide consensus on the value of a document. Furthermore, the words of different authors identify a variety of ways in which a document may be described. As a result, reference supports the indexing needs imposed by a diverse search vocabulary. To substantiate this claim, we compared the indices drawn from content and reference looking for distinct words used either alone or with other words to identify the same idea. We looked at only the root of words so that different forms of the same word were not considered distinct identifiers. In addition, all words used in the same group to identify an idea were treated as a single identifier.

For topical features, the average number of distinct identifiers per document originating in reference was 16.2, while the content indices contained an average 10.5. The mean paired difference was 5.7 with a standard deviation of 3.1 and a 90% confidence interval of $\pm$ 1.

For meta-information, the average number of distinct identifiers originating in content was 0.87 -- many sets of content indices contained none at all. The average number originating in reference was 2.5. The mean paired difference was 1.6 with a standard deviation of 1.7 and a 90% confidence interval of $\pm$ 0.6.

SUMMARY

Comparing reference to content we measured the value of each source as a basis for indexing against the metrics of topical precision, coverage for meta-information, and term diversity. By all three measures, reference demonstrated a significant advantage over content as a source of document identifiers.

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 Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of WWW '98*. Brisbane Australia, Apr 1998.

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

[4] 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.

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

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

[7] Spink, A., D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *JASIS*, 53(2): 226-234. 2001.