The Hypertext Concordance: A Better Back-of-the-Book Index

Hinrich Schütze
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304, USA
ftp://parcftp.xerox.com/pub/qca/schuetze.html

Abstract

This paper describes a tool that creates a usable index for organizing and hyperlinking related web pages. This type of hypertext construction is an important application of terminology extraction. The tool is based on a new type of hypertext: the hypertext concordance (HC). An HC is a hypertext back-of-the-book index that lists terms with context the same way a concordance does. An HC has the following characteristics: 1. Index terms are selected using a terminology extraction algorithm. 2. Occurrences of indexed terms in documents are hyperlinked to the index. 3. Term references in the index are listed with surrounding context in the manner of a concordance. 4. Each term reference in the index is hyperlinked to its document occurrence (thereby making the page numbers of paper indices unnecessary). The HC is designed for look-up (the traditional use of a book index), but also for browsing.

1 Introduction

The power of large hypertext systems such as the World Wide Web derives from hypertext markup. It enables people to find quickly what they are looking for and navigate efficiently through large volumes of information.

However, many documents and collections are not marked up with hypertext. The author may have had too little time to put in links. The documents may have been written as non-hypertext documents (ascii, word, latex). Or the documents may have been generated on the fly, e.g., as the result of a search.

This paper introduces a system that addresses the problems of unlinked or sparsely linked documents by linking them using a set of automatically extracted content words, the index terms. All occurrences of the index terms are compiled into a *hypertext concordance* (or *HC*), an index that resembles a back-of-the-book index. It has three additional features: 1. References to terms in the back-of-the-book index are given with context (in the manner of a concordance), so that the user can immediately see which of the listed references is the one of interest. 2. References to terms in the HC index are represented as hyperlinks. The user clicks on a reference to get to the corresponding document occurrence. There is no need to laboriously find a numbered page. 3. The original documents are transformed into HC documents in which the indexed terms are represented as hyperlinks. The user simply clicks on a term in a document to look up its entry in the index (e.g., to find its definition on a different page).

2 Terminology Extraction

We use Damerau's method to select index terms automatically. It compares the relative frequency in the documents of interest (coll_1) with the relative frequency in a reference collection (coll_2):

$$score(\text{word}) = \frac{f(\text{word}, \text{coll}_1)}{f(\text{coll}_1)} \div \frac{f(\text{word}, \text{coll}_2)}{f(\text{coll}_2)}$$

where $f(\text{word}_j, \text{coll}_i)$ is the frequency of word j in collection i, and $f(\text{coll}_i)$ is the total number of words in collection i. This method amounts to scoring words according to the likelihood ratio of occurring in the two collections (the documents of interest and the reference collection). This is because frequency of occurrence divided by total number of tokens in a collection is the maximum likelihood estimate of the probability of occurrence of a word.

We use the New York Times from 1990 as our reference corpus. The idea is that the New York

Times is a sufficiently broad corpus to model the frequencies of words in general English well. If a term occurs significantly more often in the documents of interest than in the New York Times, then we assume that it is an interesting phrase or technical term which should be indexed.

Terms are ranked according to the likelihood ratio and the top n terms are used as index terms. Both single words and bigrams are considered terms.

There are many other term selection algorithms. 1. Selection according to raw frequency within a text, sometimes with part-of-speech prefilters or postfilters. (Ross and Tukey, 1975; Justeson and Katz, 1995; Kupiec et al., 1995) 2. Selection according to low variance in relative position (that is, if two words occur at a fixed distance from each other whenever they co-occur, they are likely to be terminology). (Smadja, 1993) 3. Hypothesis testing and mutual information. (Church and Hanks, 1989) 4. Likelihood ratios within a text assuming a binomial distribution. (Dunning, 1993) Residual idf assuming a Poisson distribution. (Church, 1995) 5. Finding index terms based on finite-state automaton parsing. (Grefenstette, 1994) 6. Finding index terms based on full parsing. (Bourigault, 1993; Jacquemin, 1994; Strzalkowski, 1995) 7. Terminology extraction based on deeper semantic analysis and theory. (Pustejovsky et al., 1993) 8. Identifying boundaries of terminological phrases as points where the uncertainty of predicting the previous or next word is highest. (Nakagawa and Mori, 1998)

The advantage of Damerau's algorithm is its simplicity and its use of an additional source of information instead of just looking at the documents of interest. This additional source of information (the reference corpus) tells us which terms are common in the language as a whole and are therefore less likely to be of interest as index terms.

Another advantage of Damerau's algorithm is that we can apply it to very short documents (assuming we have a reference corpus). Many of the above mentioned methods require large collections to be effective.

It is probably desirable to combine different sources of information (see Evans and Zhai (1996) for one attempt). We are planning to add a part-of-speech postfilter in a manner similar to Justeson and Katz.

3 System Description

The system first asks the user to enter one URL or a list of URLs that are to be indexed. Then it applies Damerau's algorithm to extract good index terms. All words and word pairs (bigrams) are considered (except for a stop list). The top 200 terms according to the likelihood ratio are selected as index terms. The reference corpus is the 1990 New York Times.

The system then generates one or more back-of-the-book index pages (depending on the size of the index) and an HC document for each original document. The HC document is a "hyperlinked" version of the original document in which each index term occurrence has been made a hyperlink to the entry of the term in the index. Similarly, the index pages are hyperlinked to the HC documents with each reference linking back to the occurrence corresponding to it in one of the documents. The back-of-the-book index is ordered alphabetically. In the beginning, it has a list of all index terms covered, which are hyperlinked to the full entry in the index. A term's entry lists all references, one reference for each occurrence of the term in a document. Each term reference provides ten words of context, five to the left, five to the right. An identifier at the beginning of the reference line indicates the document that contains the corresponding occurrence. A word's entry in the index is hyperlinked to all entries of pairs in the index containing it and vice versa.

The user can then use a standard browser to browse the transformed document collection, either starting from the index or from the transformed documents. Filenames of the HC documents are generated by inserting a randomly generated string before the extension of the original file.

Runtime of the unoptimized current system is on the order of 10 minutes on a sparc10 for half a megabyte of text, mainly due to the complexities of determining the most frequent bigrams.

4 Related Work

This section discusses related work on automatic hypertext construction.

SuperBook. SuperBook makes search terms hyperlinks. Clicking on a search term jumps the user to the next occurrence of the term in the same document (Egan et al., 1989). In addition, SuperBook has links from the Table of Contents of a document to the corresponding sections.

SuperBook does not provide links from a document to a back-of-the-book index and links from a back-of-the-book index to documents. It does not automatically identify useful terms for an index.

Work on hyperlinking text chunks. Most work on automatic hypertext construction creates links between text chunks, based for example on similarity of vectors of keywords ((Salton et al., 1994), see also (Andersen et al., 1989)). Crouch et al. (1989) construct hyperlinks on the *document* level by linking document clusters.

Smeaton and Morissey (1995) use compactness of the hyperlink graph as an optimization criterion for choosing which text chunks to link.

Terminology Dictionaries. More closely related is work that takes technical terms from a terminology dictionary and generates hypertext from them. This is expected to give better results than an automatically extracted term list, but in many cases a hand-compiled list is not available. (Kurohashi et al., 1992)

Semantic Approaches. Two lines of work that attempt a more sophisticated semantic analysis of the domain are (Basili et al., 1995) and (Green, 1996). Generally, more expensive lexical resources and knowledge bases are required here which may not be available for a domain.

Interactive hyperlink construction. Golovchinksy reinterprets the action of following a hyperlink as refining a query: Clicking on a term roughly corresponds to adding that term to the current query. Terms to be hyperlinked are selected based on their relevance for the current working set of documents. Both term selection strategy and the entities that are linked differ from the HC. (Golovchinsky, 1997)

Approaches that use an existing hypertext structure for improving information retrieval search include (Croft and Turtle, 1989) and (Frei and Stieger, 1995).

Phrase Browsing. Nevill-Manning et al. (1997) describe the Sequitur algorithm for building a hierarchical phrase lexicon from a document collection. The user can browse the phrase lexicon to find terms that are used in the collection. The context of the terms is provided using a concordance tool.

The authors do not generate links from a back-of-the-book index to documents and from documents to a back-of-the-book index. Sequitur is designed for phrase browsing, whereas the HC is designed for document browsing.

CURL. (Unpublished work by Andrew Davison. See www.cs.mu.oz.au/ ad/ curl/ announce.html) CURL generates a hypertext structure from a specification of the logical structure of a document collection. It does not create a back-of-the-book index, nor links between documents and an index.

5 Future Work

Future work will address some remaining technical issues such as building a good interface for users to select the documents to be indexed, a principled way of determining the number of terms in the index, the selection of the reference corpus in cases where several reference corpora are available, and incorporating a morphological analyzer.

However, the main issue that needs to be addressed is evaluation. A less demanding evaluation would be to determine whether users like an HC of a document collection that has no other hypertext markup. This seems likely since if there is no alternative to finding information, then any available facility for searching and browsing should be better than having to go through documents one by one.

A better evaluation is to compare the HC with its two obvious competitors: a paper back-of-the-book index and an IR search system. Ideally, performance on a variety of tasks should be evaluated such as browsing, lookup, question answering and writing an essay (cf. (Egan et al., 1989)). Note that even if the HC were found to perform less well than these alternatives, there could still be an argument for its use in situations where paper access is not feasible or limited resources preclude running a search engine.

It would be desirable to have separate evaluations of the terminology extraction method used (automatic or manual) and the architecture of the HC. A professional indexer will likely generate better index terms than an automatic system. But this manually generated list, when made the basis of an HC, could still be easier to use than the corresponding paper version.

References

M. H. Andersen, J. Nielsen, and H. Rasmussen. 1989. A similarity-based hypertext browser for reading the UNIX network news. *Hypermedia*, 1(3):255–265.

Roberto Basili, Fabrizio Grisoli, and Maria Teresa Pazienza. 1995. Hirma: Hypertextual information retrieval system managed by ariosto. *Data and Knowledge Engineering*, 17(3):187–213.

Didier Bourigault. 1993. An endogeneous corpus-based method for structural noun phrase disambiguation. In *Proc. of EACL*, pages 81–86.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *Proc. of ACL*, pages 76–83.

Kenneth Ward Church. 1995. One term or two? In *Proc. of SIGIR*, pages 310–318.

W. Bruce Croft and Howard Turtle. 1989. A retrieval model for incorporating hypertext links. In *Proc. of Hypertext*, pages 213–224.

Donald B. Crouch, Carolyn J. Crouch, and Glenn Andreas. 1989. The use of cluster hierarchies in hypertext information retrieval. In *Proc. of Hypertext*, pages 225–237.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comp. Ling.*, 19(1):61–74.

Dennis E. Egan, Joel R. Remde, Louis M. Gomez, Thomas K. Landauer, Jennifer Eberhardt, and Carol C. Lochbaum. 1989. Formative design-evaluation of superbook. *ACM TOIS*, 7(1):30–57.

David A. Evans and Chengxiang Zhai. 1996. Noun-phrase analysis in unrestricted text for information retrieval. In *Proc. of ACL*, pages 17–24.

H.P. Frei and D. Stieger. 1995. The use of semantic links in hypertext information retrieval. *IP&M*, 31(1):1–13.

Gene Golovchinsky. 1997. Queries? Links? Is there a difference? In *Proc. of CHI 97*, pages 407–414.

Stephen J. Green. 1996. Using lexical chains to build hypertext links in newspaper articles. In *AAAI 96 Workshop on Internet-based Information Systems.*

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Press, Boston.

Christian Jacquemin. 1994. FASTR: A unification-based front-end to automatic indexing. In *Proc. of RIAO*, pages 34–47.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proc. of SIGIR*, pages 68–73.

Sadao Kurohashi, Makoto Nagao, Satoshi Sato, and Masahiko Murakami. 1992. A method of automatic hypertext construction from an encyclopedic dictionary of a specific field. In *Proc. of ANLP 3*, pages 239–240.

Hiroshi Nakagawa and Tatsunori Mori. 1998. Nested collocation and compound noun for term extraction. In *these proceedings.*

Craig G. Nevill-Manning, Ian H. Witten, and Gordon W. Paynter. 1997. Browsing in digital libraries: A phrase-based approach. In *Proc. of ACM DigLib*, pages 230–236.

James Pustejovsky, Sabine Bergler, and Peter Anick. 1993. Lexical semantic techniques for corpus analysis. *Comp. Ling.*, 19(2):331–358.

Ian C. Ross and John W. Tukey. 1975. Introduction to these volumes. In John Wilder Tukey, editor, *Index to Statistics and Probability*, pages iv–x. R&D Press, Los Altos CA.

G. Salton, J. Allan, C. Buckley, and A. Singhal. 1994. Automatic analysis, theme generation and summarization of machine-readable texts. *Science*, 264:1421–1426.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Comp. Ling.*, 19(1):143–177.

Alan F. Smeaton and Patrick J. Morissey. 1995. Experiments on the automatic construction of hypertext from texts. *The New Review of Hypermedia and Multimedia*, 1:23–39.

Tomek Strzalkowski. 1995. Natural language information retrieval. *IP&M*, 31(3):397–417.