Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2002 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2002

SUMMARIZING SEARCH RESULTS WITH AUTOMATIC TABLES OF CONTENTS

Yi-Fang Wu
New Jersey Institute of Technology

Chatchai Rakthin
New Jersey Institute of Technology

Changzhi Li
New Jersey Institute of Technology

SUMMARIZING SEARCH RESULTS WITH AUTOMATIC TABLES OF CONTENTS

Yi-Fang Brook Wu, Chatchai Rakthin, and Changzhi Li
Information Systems Department
New Jersey Institute of Technology
wu@njit.edu crakthin@hotmail.com
cl6@njit.edu

Abstract

Methods to help efficiently digest and use information have become an important research area in resolving information overload. While automatic text summarization research demonstrates significant efforts in this regard, the processes are complex and results have poorer quality than those created manually by experts. The generated summary tends to be rigid as the systems are designed for specific domains and applications. Our Automatic Table of Content Developer system, based on the Probability of Co-occurrence Analysis (POCA) technique, creates outputs that are content dependent, and yet the system can adapt to all document domains. We thus propose the resultant table-of-contents (TOC) display as an alternative to help users digest a large document collection using less time and effort.

Keywords: Summarization, co-occurrence analysis, concept hierarchies, automatic table-of-contents

Introduction

According to the study of Kleijnen and Groenendaal (2000), about 170 online peer-review journals and conference sites have been established to support work in the IS related area. Although the quality requirements of accepted work limit the number of approved publications, tremendous amounts of documents are still continuously being created. Writing research papers is tedious and error-prone as a result of the voluminous data the authors need to process. Research in the area of automatic text summarization can help people summarize the data by providing an overview of documents.

Our system classifies closely related terms, according to user queries, and presents them in a hierarchical structure similar to a TOC of a book. Kolcz et al. (2001) suggest that text categorization in a hierarchical or tree-like structure is useful in helping users retrieve or focus on documents at different levels of relevancy. This organization format might be a solution to "sequentiality" effect discussed in (Korfhage, 1997). The theory is that most search systems list returned documents in linear order, but users are satisfied if the relevant results are shown first. By organizing search results in the TOC style and providing a link directly from TOC structure to related topics in the text, users can quickly browse returned documents without reading the full text. Thus, the "sequentiality" effect is expected to be lessened. This TOC display can also serve the purpose of the summary of search results, because all important and relevant concepts are listed in the TOC.

Text Summarization

The automatic text summarization is a computerized process for preparing documents, extracting important parts, and presenting a short, well-organized summary that meets a particular user's need and task (Mani 2001). The major areas of automatic text summarization include paragraph-based, sentence-based, and discourse model-based summarization (Amitay and Paris 2000). While the paragraph-based technique is useful for long documents containing several important topics, the other two techniques focus on a finer level of extraction and should produce a summary which best represents the document's content.

While early text summarization research used a statistical approach to summarize text, later research focuses more on the semantic understanding of the document. Luhn's (1968) statistical approach was based on the intuition that the most frequently occurring words or phrases expose the most important concepts in documents; hence, he combined sentences containing those terms into a summary. However, Silber and McCoy (2000) argued that this statistical approach may give results that are not relevant to the content. Morris and Hirst (1991) focused their work on the semantic aspect of documents using the lexical chains analysis. They believed that strong lexical chains (the related terms that best represent documents) can be used to generate a good summary.

Chuang and Yang (2000) found that automatic text summarization requires a good understanding of the text. They proposed to create a summary from extracts, e.g. noun phrases or sentences, of the document. Their work was based on a tree structure called the "Rhetorical Structure Theory" (RST), which breaks a complex sentence into clauses: the more important part or nucleus will be placed on top of the tree and the subordinate parts or satellites will be located deeper on the tree. This idea demonstrates that there are some relationships between general and specific concepts which are helpful to distinguish important parts of the text in order to create summaries that are independent of the domain of documents and users. However, due to the different dimensions in texts such as writing styles, word usages, and lengths, it is difficult to design an automatic summarization system that is comparable to the capability of a domain expert.

Text Classification and TOC Structure

The TOC style we are proposing is similar to "term trees." This structure, according to Rieh and Xie (2001), might improve user satisfaction and retrieval effectiveness because of its ability to facilitate query refinement and reformulation. Unfortunately, all major search engines lack this important feature. Toutanova et al. (2001) focus their work on classifying text into classes or topics that can describe the documents. Their study found that placing more general terms on higher nodes and more specific terms on lower levels should help users in locating their desired information more easily. Current developments include either manual, e.g. thesauri, or automatic methods, e.g. results based on our POCA technique. The former generate semantically related terms, while the latter develop "functionally related terms," which according to Buckland (2001), are words that are highly related and are helpful in retrieval.

While previous research proposed to develop only basic concept hierarchies (Sanderson and Croft 1999), our study offers an extended version of concept hierarchies that is based on POCA (Wu 2000). This automatic TOC improves basic concept hierarchies by adding a function to point to relevant portions of text in the retrieved documents. The hierarchical structure of our automatic TOC is similar to a familiar TOC in a book which lists important topics and their locations (page numbers) while organizing topics from broader ones (chapters) to narrower ones (sections). The structure is helpful because people are used to searching for things that are organized in hierarchical formats (Chen et al. 1998). In addition, our automatic TOC can work like multiple searches within results, meaning that it allows users to browse the resulting concept hierarchies for a broader topic and locate a narrower subtopic among all relevant documents that already contain the broader topic. In other words, a user can find, within a set of returned documents, a topic within a topic within a topic.

The Probability of Co-occurrence Analysis

Sanderson and Croft (1999) propose using subsumption to develop concept hierarchies. Their theory is as follows. Suppose X and Y are two different terms in a corpus. If Y appears in a subset of documents in which X appears, then X is said to subsume Y. Therefore, because X subsumes Y and because X is more frequent in the hierarchy, X is the parent of Y (Figure 1). According to Sanderson and Croft's theory, closely related concepts are likely to appear in the same text, and broader (general) concepts usually appear more frequently in the text. Thus, by analyzing co-occurrences between term pairs, the relationship between the broader and the narrower terms can be determined.

They then modified the rule to include imperfect subsumption pairs, in which there are only a few occurrences of Y that are not accompanied by X (Figure 2).

The Probability of Co-occurrence Analysis (POCA) (Wu 2000) was developed based on the subsumption theory and conditional probability and is re-defined as follows:

$$P(X|Y) > P(Y|X),\ P(X|Y) >= N,\ \text{where } 0 < N < 1$$

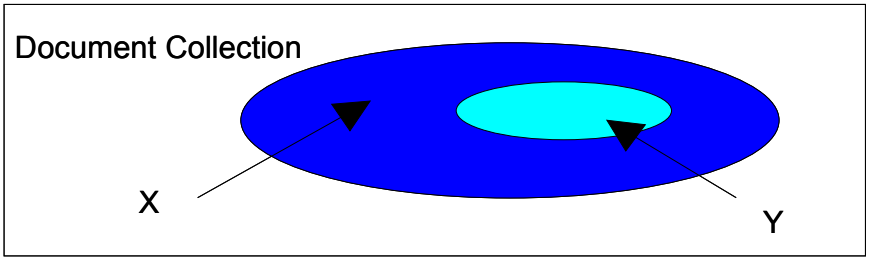

Figure 1. A Demonstration of an Example of Subsumption

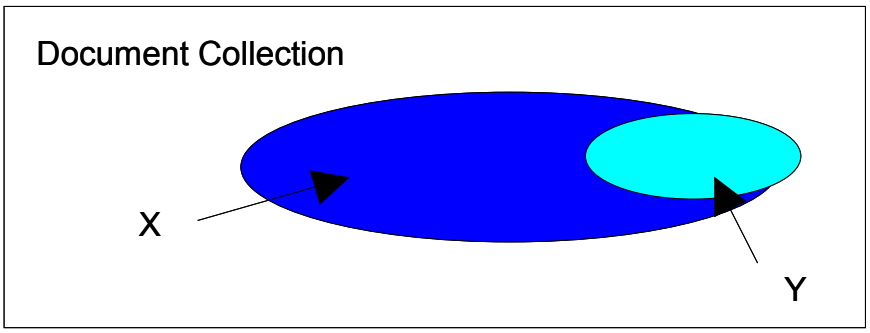

Figure 2. A Demonstration of an Imperfect Subsumption

If a term pair (X, Y) fulfills the above set of inequalities, X occurs more frequently than Y in the document collection, and is therefore broader than Y. Note, the threshold, N, affects the number of term pairs derived; namely, larger N results in a smaller number of term pairs.

Automatic TOC Developer

Our program has lexical processing, natural language processing, and concept hierarchies building capabilities. To develop concept hierarchies, the program searches for the best matching paragraphs which contain the query terms, and treats them as separate documents. Then, noun phrases are collected via the Noun Phrase Extractor program we developed, which performs morphological and syntactic analysis and is based on the lexical databases of the Wordnet (Miller 1995). Only noun phrases are collected and automatically indexed because they are more likely to be the concepts people express in writing. While a long document can contain several different topics, the wanted query terms and related concepts are more likely to be present together in the same paragraphs. Therefore, we believe that collecting best matching passages for analysis should create the TOC that represents closely related and thus useful terms.

The system allows users to choose different settings: indexing styles, e.g. automatic indexing and query matching in paragraph or document level, threshold values (default: 0.8), and term levels (default: noun phrase). For indexing styles, automatic indexing is used to index all noun phrases; query matching is used to index noun phrases in best matching passages. After the program finishes the indexing and concept hierarchy developing, it will show the number of retrieved documents as well as the number of noun phrases. Users are allowed to choose different thresholds. Based on our observations, the threshold amount of 0.8 yields the best results, as reported in Sanderson and Croft (1999)'s study.

In this study, we ran the Automatic TOC Developer program on a collection of 3,138 abstracts collected from Journal of Management Information Systems, Management Science, Communications of the ACM, MIS Quarterly, and Information Systems Research. These papers were published between 1988-1998. The query "data mining" was used to perform a search. The result includes 13 retrieved documents and 91 extracted noun phrases. The left window displays the TOC structure of returned documents, while the right one shows the related document that the selected term refers to (Figure 3). The bottom window demonstrates the information about the selected term "knowledge discovery applications."

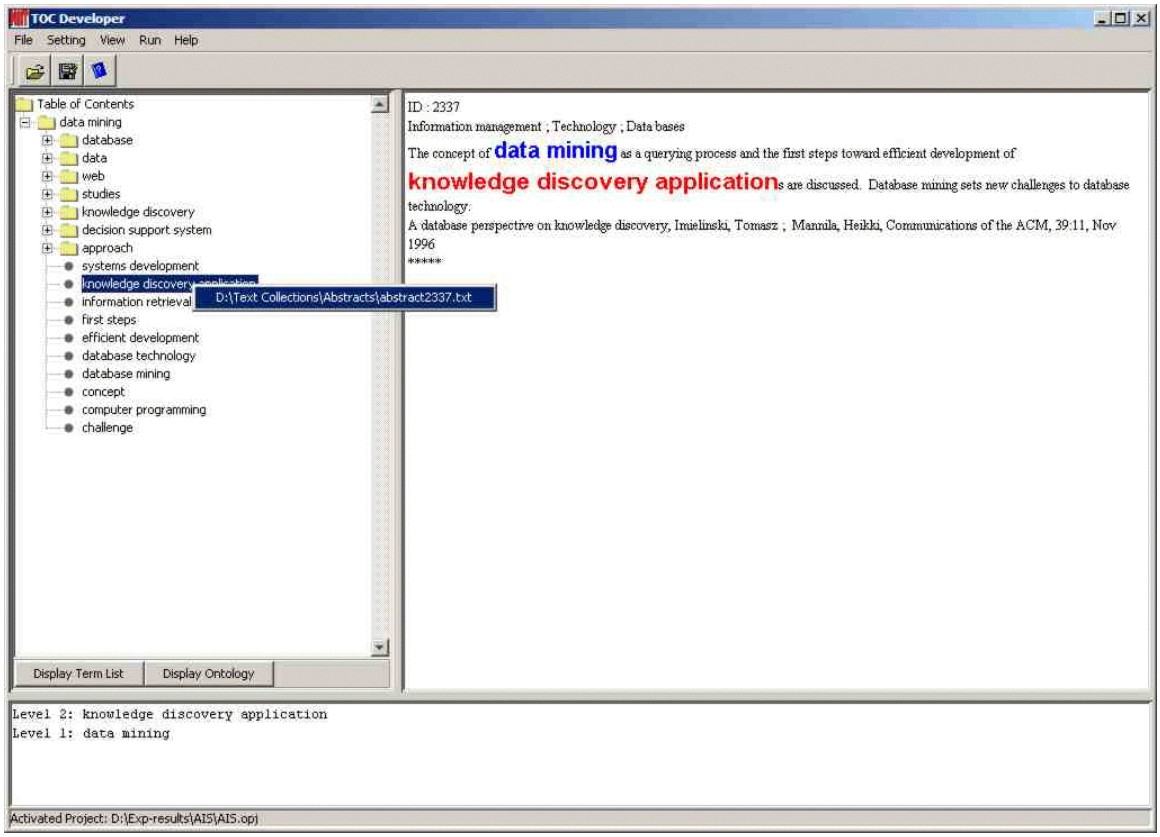

Figure 3. System Output Showing TOC on the Left-Hand Side

The TOC structure demonstrates the parent-child relationship, in which the higher term in the same level more frequently co-occurs with its parent. Placing the mouse pointer on any term will show the location of documents containing that term. In addition, clicking the mouse on a term allows users to see information regarding its parent term, the number of its child terms, and the number of relevant documents. The system allows switching the mode between viewing the full text or the relevant sentences. This feature is useful in locating relevant concepts faster in a long document. In the document, the keywords of the selected concept hierarchy branch, i.e. "data mining" → "knowledge discovery application" in this example, are enlarged and highlighted in color corresponding to their term levels.

Discussions and Future Work

Even though we only performed the small scale test run which only generated concept hierarchies for 13 returned abstracts, we expect the benefits will be more obvious when a larger text collection of longer documents is used. This is more reflective of the large amounts of text people have to deal with every day.

The results of the automatic TOC developer are better than those of traditional document classification techniques using polythetic classification methods, where each document can be assigned to only one category and might be a significant limitation for classifying long documents containing several topics. Automatic TOC is useful as an alternative to text summarization by allowing users to spend less time and effort to digest large documents. The classification structure of our automatic TOC should be useful in human-computer interaction research.

The next stage of this research will focus on evaluation of how automatic TOC might help users to find relevant information in documents. We plan to design a user study similar to Chen et al. (1998)'s experiment. Our system will use the search results generated by Google as returned documents for TOC development. The query used to retrieve documents will be the same, so identical search results will be obtained, with the only difference being the presentation format of returned documents. One group of subjects will perform their browsing task using Google, while another group will use the TOC display. We will make a

comparison based on the time the subjects use to find a certain number of relevant documents and how they navigate returned document sets. We will also evaluate the user satisfaction toward the system. This experiment will be conducted in Fall 2002.

References

Amitay, E. and Paris, C. *Automatically Summarising Web Sites - Is There A Way Around It?*, Proceedings of the 9th International Conference on Information and Knowledge Management, Virginia, November 2000, pp. 173-179.

Buckland, M. *Entry Vocabulary, Intermediaries, and Retrieval Performance*, Proceedings of ASIS&T Annual Meeting, Washington DC, November 2001, pp. 112-117.

Chen, H., Houston, A. L., Sewell, R. R., and Schatz, B. R. *Internet Browsing and Searching: User Evaluation of Category Map and Concept Space Techniques*, JASIS, 49(7), 1998, pp. 582-603.

Chuang, W. T. and Yang, J. *Extracting Sentence Segments for Text Summarization: A Machine Learning Approach*, Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Greece, July 2000, pp. 152-159.

Kleijnen, J. P. C. and Groenendaal, W. V. *Measuring the quality of publications: new methodology and case study*, Information Processing & Management, 36(4), 2000, pp. 551-570.

Kolcz, A., Prabakarmurthi, V., Kalita, J. K. *Summarization as Feature Selection for Text Categorization*, Proceedings of the 10th International Conference on Conference on Information and Knowledge Management, Georgia, October 2001, pp. 365-370.

Korfhage, R. R. *Information Storage and Retrieval*, New York: John Wiley and Sons, Inc, 1997.

Luhn, H. P. *The automatic creation of literature abstracts*. In Schultz, editor, H. P. Luhn: Pioneer of Information Science, Spartan, 1968.

Mani, I. *Recent Developments in Text Summarization*, Proceedings of the 10th International Conference on Conference on Information and Knowledge Management, Georgia, October 2001, pp. 529-531.

Miller, G. A. *WordNet: A Lexical Database for English*, Communications of the ACM, 38(11), November 1995, pp. 39-41.

Morris, J. and Hirst, G. *Lexical cohesion computed by thesaural relations as an indicator of the structure of the text*, In Computational Linguistics, 18(1), 1991, pp. 21-45.

Rieh, S. Y. and Xie, H. *Patterns and Sequences of Multiple Query Reformulations*, Proceedings of ASIS&T Annual Meeting, Washington DC, November 2001, pp. 246-255.

Sanderson, M. and Croft, B. *Deriving concept hierarchies from text*, Proceedings on the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 206-213.

Silber, G. H. and McCoy, K. F. *Efficient Text Summarization Using Lexical Chains*, Proceedings of the 2000 International Conference on Intelligent User Interfaces, Louisiana, January 2000, pp. 252-255.

Toutanova, K., Chen, F., Popat, K. and Hofmann, T. *Text Classification in a Hierarchical Mixture Model for Small Training Sets*, Proceedings of the 10th International Conference on Conference on Information and Knowledge Management, Georgia, October 2001, pp. 105-113.

Wu, Y. -f. *Automatic Concept Organization: Organizing Concepts from Text Through Probability of Co-occurrence Analysis*, In Proceedings of the 11th ASIST SIG/CR Classification Research Workshop, Chicago, November 2000.