PROSIDING

SEMINAR KEBANGSAAN

SAINS, TEKNOLOGI & SAINS SOSIAL

27 ~ 28 MEI 2002
HOTEL VISTANA, KUANTAN, PAHANG

Anjuran :

Universiti Teknologi MARA
Cawangan Pahang

Dengan Kerjasama

Kerajaan
Negeri Pahang Darul Makmur

JILID 2

AUTOMATIC CATEGORIZATION OF BOOK COLLECTIONS

N. IDRIS AND A. DENNIS
Department of Artificial Intelligence, Faculty of Computer Science and Information Technology,
University of Malaya, Kuala Lumpur 50603, Malaysia.

Abstract

Automatic text categorization is an important research area and has a potential for many text-based applications. This paper discussed the role of the information retrieval (IR) as a way of categorizing books automatically called Smart Database Assistant. It is a system for accessing and categorizing collections of books, which involved two modules, the Public User Module and the Administrator Module. These two modules offer an approach to accessing and categorizing text-based books collections by the keywords of the contents or themes that are described by a user which based on the indexing process. To test the effectiveness of the developed system, an experiment was conducted against a number of books from different category. The result showed that the information retrieval offers an acceptable performance where it is applicable for categorizing the book collections.

Keyword: Information retrieval, indexing process, stemming, categorization

INTRODUCTION

Automatic categorization is an important and powerful tool for supporting electronic records management. Automatic text categorization is a process of assigning individual text units automatically into some predefined categories. It requires flexibility to handle a large volume of data efficiently such as to process and to understand the content of the data to a degree that will give meaningful results. It is also an important research area and has potential for many text-based applications, which consists of the processing of a sequence of texts like books, newspapers, reports and other sources that is based on texts.

Some usages of text categorization have been: to assign subject categorization to document in the support to text retrieval and library organization [8], or to aid the human assignment of such categories, such as categorizing articles in magazine [7], categorizing stories by activities [5] and categorizing broadcasting news [6]. Categorization is a problem that dealt with so many applications and each application has its own purpose in using the technique. In the process of categorization of electronic documents, categories are typically used as a means of organizing and getting information in a collection of documents. In other application such as the categorization of articles in magazine, the technique is used to classify a variety of articles of different magazines and columns.

This paper describes our effort to provide an automatic categorization of book collections, which classify variety of books of different categories. In the second section, we reviewed the materials and methods used in this project. Then, we described the process of categorizing the books with the experiments and concluded with a discussion of the performance of the system.

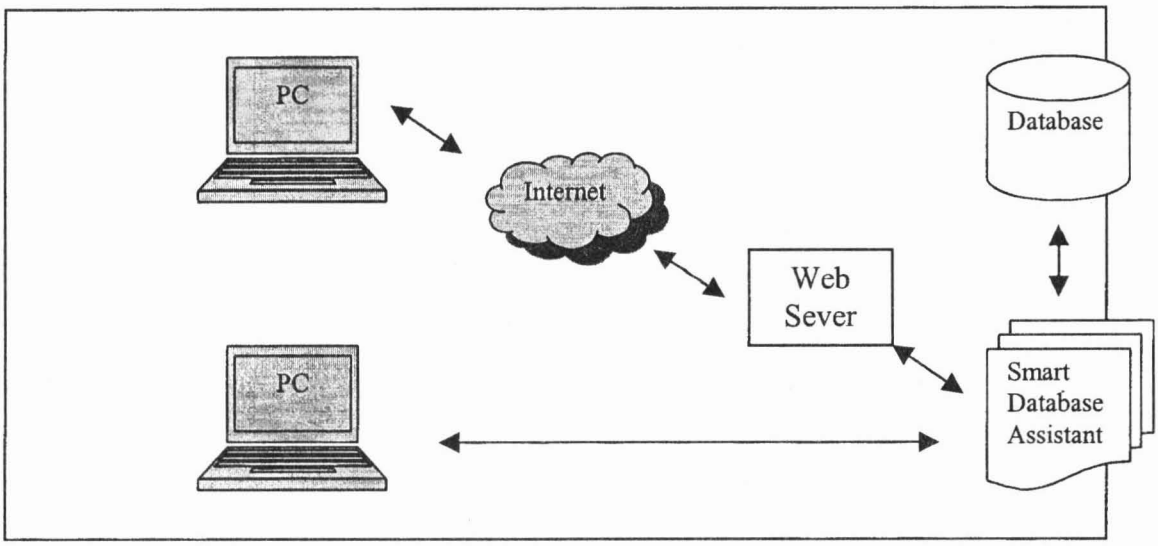
MATERIALS AND METHODS

There are many techniques that can be applied to the problems in the categorization of book collections. The major focus in this project is on the information retrieval, which concerned with selecting information from a collection that may be of interest to a searcher [9]. It is an established technology, which has been delivering solutions to the users for more than three decades and still an active research area [2]. It also has gained attention since in the 1940's [10].

Figure 1 showed the general view of the system where it contains major components of the system and shows the information flows between them. The concept of the system is based on the integration of Java and Prolog, which combines the advantages of each language, Java for the Object Oriented programming and Prolog for the powerful inference engine. In this system, Prolog will provide the Smart Database Assistant Engine and the knowledge base for storing all the set of rules that the engine should abide. Prolog

will be accessed by Java, which receive input from the user, handle the text processing and has direct access to the database.

Figure 1. The architecture of the system.



The Smart Database Assistant is divided into two main modules, the Public User Module and the Administrator Module. The Public User Module is where the user or public user can access the book collections. The module is limited to only retrieving and reviewing the records. The main function of this module lies on the ability of the query component. Since the query is based on the natural language, the components interact directly with the engine, which processes the text input and is called indexing. Indexing is an important process in an information retrieval where it organizes text documents based on the contents [1]. It is performed by assigning each document with keywords representing the document and the keywords must reflect the content of the document to allow effective keyword searching. The main purpose of the indexing is to generate transactions whereby the words or terms which are significant will be identified.

When doing the query searching, the Public User Module will send the query to the engine. The engine then parsed the text and removed all the common words through a process called stopwords removal. Stopwords are those words that are frequently used and have little information value [3]. Stopwords are important because if a word appears many times in a document, it is less useful as a key to that document than the words that occur only a few times. As an example, the word 'music' may occur once or twice in the text, but the word 'is' may occur several times in the same text. Similarly, the word 'is' will also occur with high frequency in all other texts. Due to this situation, the stopwords are removed from the documents during the indexing process and consequently enhance the speed of the indexing process.

After removing all the stopwords from the text, the remaining words will be stemmed by a process called words stemming, to produce the keywords for the search. Stemming can be defined as a process of extracting each word from text document, reducing it to a probable root word [4]. It is a technique of linguistic normalization, in which the variant forms of a word are reduced to a common form. There is no doubt that stemmed word are more effectively used than the ordinary word in performing the information retrieval task. By reducing the morphological variance of terms, we can improve the query matching process. A stem word is produced by removing affixes from words the text document. Affix is the verbal elements that is attached to the word whether at the beginning of the word (prefix), at the end of the word (suffix) or at the middle of the word (infix). However, the stemmer for this system is only concerned with the removal of suffixes and it has been found to be sufficient for the English words. For example, an English word 'music' has similar meaning to the words such as 'musical' and 'musician'. The engine then analyzed the database for any matching words or words that is relevant to the keywords. When the result was found and weighted, the engine passed the result to the module to be displayed as a hit list.

All books available in the database were kept according to their category such as Fictions, Mathematic, Computing and Mythology. This system offers an automatic categorizing engine to help administrator to classify new books based on the descriptions or summary of the books. It is one of the components of the Administrator Module and the main function of the module. The module called the engine and passed the text describing the theme of the book. The text is stemmed, processed and analyzed to find the most relevant matching category. After the result was found, the engine will confirm the result by sending the result back to the module to be displayed before it can be saved in the database.

DISCUSSION

This paper represents the work on accessing and categorizing collections of books using the information retrieval concept. The system although does not have powerful features to some extent, still it has some strength of its own when compared to some existing database search engine. It provides user a natural language query and display only the relevant information that they want, compare to conventional search engine that will flood user with irrelevant information. The system also provides a tool to help user decide which category should the following book goes to base on the book's description. This is very useful because the users might want to organize their database more efficiently. However, there are many directions in which this system could be enhanced. For example, the query parser engine could be converted to other language so that besides English, other language could also be used to query for books. This will be our future work in order to produce a smart database system which capable of managing the book collections efficiently.

REFERENCES

[1] Adriani, M. and Croft W. B. 1997. Retrieval Effectiveness of Various Indexing Techniques on Indonesian News Articles.

[2] Agosti, M. and Smeaton, A. F. 1996. *Information Retrieval and Hypertext.* Kluwer Academic Publishers.

[3] Callan, J. P.; Croft W. B. and Broglio, J. 1995. TREC and TIPSTER Experiments with INQUERY. *Information Processing and Management*, pages 327-343.

[4] Fuller, M. and Zobel, J. 1998. Conflation-based Comparison of Stemming Algorithms. *In Proceedings of the Third Australian Document Computing Symposium (ADCS'98).*

[5] Gordon, A. S. 2000. Accessing Story Collections by Activities. Technical report, IBM Thomas J. Watson Research Center.

[6] Luo, Huitao. 1999. Experiments on Automatic Categorization of Broadcasting News. Technical report, AT&T Lab.

[7] Moens, M. F. and Dumortier, J. 1999. Automatic Categorization of Magazine Articles. *In Proceedings Informatiewetenschap 1999.*

[8] Rajashekar, T. B. and Croft, W. B. 1995. Combining Automatic and Manual Index Representations in Probabilistic Retrieval. *Journal of the American Society for Information Science*, 46(4):272-283.

[9] Turtle, H. R. 1991. Inference Networks for Document Retrieval. PhD thesis, University of Massachusetts.

[10] Van Rijsbergen, C. J. 1979. *Information Retrieval.* Butterworths, London, second edition.