Retrieval Effectiveness Of Various Indexing Techniques
On Indonesian News Articles

Mirna Adriani
Fakultas Ilmu Komputer
Universitas Indonesia
Depok 16424
Indonesia
mirna@cs.ui.ac.id

W. Bruce Croft
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts, Amherst
Amherst, MA 01003-4610, USA
croft@cs.umass.edu

May 1997

Abstract

The performance of various indexing techniques, namely, manual indexing, automatic indexing, and combined manual and automatic indexing in retrieving Indonesian news articles is evaluated. The result of using structured and Boolean queries show that the combined indexing technique is more effective than the other techniques. The results demonstrate that the indexing techniques and query methods which work with English texts are also applicable to Indonesian texts.

1. Introduction

Organizing text documents based on their contents, called indexing, is an important process in an information retrieval system. Basically, indexing is performed by assigning each document with keywords or descriptive terms representing the document. The assigned terms must reflect the content of the document to allow effective keyword searching. In the past, indexing has been done manually by trained persons who are familiar with the topics of the texts.

Today, with the increasing availability of electronic texts online, manual indexing is obviously too slow and, needless to mention, too expensive. Automatic text indexing which is much faster and less error-prone has become a common place. Research on English texts has shown that the retrieval effectiveness of automatic indexing is comparable to that of manual indexing [Sparck Jones, 1974; Salton, 1986].

The main goal of this study is to see whether the result of such experiments on English texts are consistent with those for other languages, in particular, Indonesian texts. We evaluate the effectiveness of automatic, manual, and combined automatic-manual indexing techniques for natural language, Boolean, and phrase queries. In the outset, we expected that the combined indexing technique would perform better for various types of queries than the other techniques. In addition, we are also interested in seeing whether phrase and Boolean queries perform better than natural language queries, as is the case for English text retrieval [Croft et.al., 1991].

The query types evaluated are natural language, phrase, and Boolean queries. A natural language query specifies the user's information need in a natural language sentence or sentences. A phrase query contains phrases representing concepts of interest to the user. A Boolean query makes use of Boolean operators such as 'and' and 'or', in addition to keywords, to specify the search criterion. In this study, we also investigate the effect of applying 'strict' operators, namely, the word-proximity operator on the phrase queries and the #band operators on the Boolean queries.

2. Previous Work

Many researchers have compared the effectiveness between manual and automatic indexing techniques. Manual indices were often presumed to be better than machine generated indices. However, it has been demonstrated that both indexing techniques are equally effective for text retrieval [Salton, 1989]. The retrieval performance were also showed positive improvement if both techniques were combined compared to individual indexing technique [Callan et al., 1993; Rajashekar & Croft, 1993].

In order to thoroughly compare the performance of the indexing techniques, many researchers tested various query formulation methods, or query types, on each of the techniques. Previous studies with English texts showed that phrase and Boolean queries perform better than natural language queries [Croft et.al., 1991; Belkin et.al., 1993]. The results of the studies also show that adding word-proximity operators to the Boolean and phrase queries, resulting in structured queries, can further improve the queries' effectiveness for large collections.

A similar study using Japanese texts showed that phrase queries improve the effectiveness of the natural language queries [Fujii & Croft, 1993]. Unlike the previous studies, this study demonstrated that applying proximity operators requiring the co-occurence of words in a document decreases the performance of the phrase queries.

A recent study by Turtle [Turtle, 1994] produced a different conclusion, in that natural language queries perform better than Boolean queries in retrieving English legal documents. Yet, another study found that both natural language and Boolean types of queries, generated by real user, are equally effective in retrieving medical related documents [Hersh & Hickam, 1995].

3. The Experiment

For the experiment, we built the text database using a version of INQUERY which has been modified to handle Indonesian text. In particular, we incorporated an Indonesian word-stemmer module into INQUERY. INQUERY is a retrieval system, based on a probabilistic retrieval model, which was developed at the Information Retrieval Laboratory, the University of Massachusetts [Callan et.al., 1992].

We conducted some experiments to investigate the effectiveness of three types of queries, natural language, phrase and Boolean queries in retrieving Indonesian texts. The phrase queries are specified using the #phrase operators which require the query words to occur together within three-words distance. The Boolean queries are specified using the #and and #or Boolean operators.

We also evaluate the effect of applying a number of constraining operators to the queries, namely, the #1 proximity operator to the phrase queries and the #band operator to the Boolean queries. The #1 proximity operator requires the co-occurrence of the query words next to each other. The #band operator increases the selectivity of the Boolean queries by requiring that all query words, and no less, must occur in a document. As a comparison, the Boolean #and operator in INQUERY means that a document receives a higher score if more query words are found in the document.

In this study, we use a collection of 19,733 Indonesian newspaper articles from *Kompas* (an Indonesian daily newspaper). An automatic index was built using INQUERY and a manual index was obtained from human indexers at the newspaper publisher's office. A manual-automatic combined index was obtained by first adding the manual index terms to the text documents, and then built the index automatically using INQUERY.

The natural language queries for our experiments were created based on the TREC short queries topics [Harman, 1995]. Query topics which are relevant to the Indonesian text documents were selected. We used 25 queries for all of the experiments. The phrase and Boolean queries were constructed manually transforming from the natural language queries.

The effectiveness of the indexing techniques was measured by their average retrieval recalls and precisions for the top-25 documents in the document rank list. The small number of documents (25) to evaluate was used because we are particularly interested in measuring the effectiveness the techniques were for quick retrieval, where the user did not want to spend time checking the relevance of too many retrieved documents.

4. The Experimental Results

The average precision as shown in Table 1. indicates that the combined indexing technique performs better than the automatic and manual indexing for the natural language and phrase queries. Among the other techniques, the automatic indexing technique shows better average performance than that of the manual indexing technique.

	Automatic Indexing	Combined Indexing	Manual Indexing
NL	0.2579	0.2669	0.1706
Phrase (#phrase)	0.2653 (2.52)	0.2755 (2.85)	0.2443 (35.87)
Boolean (#and)	0.2269 (-12.29)	0.2298 (-12.18)	0.2074 (43.20)

Table 1. Average retrieval precisions for the three basic query types for each of the three indexing techniques.

In this set of experiments, the average performance of the natural language queries is taken as the base performance with which the performance of the other query is compared.

For all of the indexing techniques, the phrase queries perform better than the natural language queries. The difference is very significant (35.87%) for the manual indexing technique, but only slightly, 2.52% and 2.85%, for the automatic and the combined indexing techniques, respectively. The main reason is that the manual index consists of mostly phrases which are commonly used by humans (journalist) for categorizing news articles, and some of these phrases happened to coincide with the phrases used in our queries.

The Boolean queries perform poorly, as compared to the natural language queries, for both the automatic indexing (-12.29%) and the combined indexing (-12.18%) techniques, but show a superior average performance for the manual indexing, i.e. (43.20%) better than the natural language indexing performance as shown in Table 1.

In another set of the experiments, we used structured queries obtained by applying #1 proximity operators to the phrase queries and #band operators to the Boolean queries.

The results, as shown in Table 2. demonstrate that applying the #1 proximity operator to the phrase queries decreases the average retrieval precision for all of the indexing techniques. Whereas, the #band operator improves the performance of the Boolean queries, except for the manual indexing technique. The main reason for this exceptional case is that the vocabulary of terms used in the manual index is much smaller than that of the query terms. In other words, there are more query terms not found in the manual index than in the automatic and the combined indices.

	Automatic Indexing	Combined Indexing	Manual Indexing
Phrase (#phrase)	0.2653	0.2755	0.2443
Phrase (#1)	0.2113 (-20.08)	0.2181 (-20.55)	0.1620 (-30.11)
Boolean (#and)	0.2269	0.2298	0.2074
Boolean (#band)	0.2550 (12.73)	0.2730 (16.47)	0.1866 (-23.62)

Table 2. Average retrieval precisions for the plain and the enhanced phrase queries, and the plain and the enhanced Boolean queries for the three indexing techniques.

The results for the phrase queries suggest that Indonesian texts, as texts in many other languages, does not use strict word ordering rules, for instance, a word can be inserted between the words in a phrase, a modifier word can be moved within a phrase, etc.

5. Conclusion

In this study, we learned that the combination of manual and automatic indexing techniques is more effective than the individual techniques, at least for Indonesian texts and English texts (as suggested by other researchers).

The results of using different types of query for Indonesian texts agree with the results for English texts in most other studies, i.e., that natural language queries are at least as effective as Boolean queries, with an exception for, in our case, the manual indexing technique. Compared to the results for Japanese texts, our results agree in that phrase queries are more effective than natural language queries, and that making the phrase queries stricter do not improve their effectiveness.

In short, there is still an open issue to be studied whether different query formulation methods can improve retrieval performance for different collections and different languages.

Acknowledgments

We would like to thank Hardanto, Andrey Andoko, and Herdanto from Kompas (http://www.kompas.com) who provide the manual indexing and the Indonesian text collection on this study.

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

Belkin, N., C. Cool, W. Bruce Croft, and J. D. Callan. The Effect of Multiple Query Representations on Information Retrieval System Performance. *In Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval,* 339-346, 1993..

Callan, J. P., W. Bruce Croft, and S. M. Harding. The Inquery Retrieval System. *3rd International Conference on Database and Expert Systems Applications*, 1992.

Callan, J. P. and W. Bruce Croft. An Evaluation of Query Processing Strategies using TIPSTER collection. *In Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval,* 347-356, 1993.

Croft, W. Bruce, Howard R. Turtle, and David D. Lewis. The Use of Phrases and Structured Queries in Information Retrieval. *In Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval,* pp. 32-45, 1991.

Fujii, Hideo and W. Bruce Croft. A Comparison of Indexing Techniques for Japanese Text Retrieval. *In Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval,* 237-246, 1993.

Harman, Donna. Overview of the Fourth Text Retrieval Conference (TREC-4). *Proceedings of the Fourth Text Retrieval Conference (TREC-4),* 1995.

Hersh, William R. and David H. Hickam. An Evaluation of Interactive Boolean and Natural Language Searching with an Online Medical Textbook. *Journal of the American Society for Information Science*: 46(7), 478-489, 1995.

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

Salton, Gerard. Another Look at Automatic Text-Retrieval Systems. *Communications of the ACM* 29(7), 648-656, July 1986.

Salton, G. *Automatic Text processing*. Addison-Wesley Publishing Company, Reading, MA: 1989.

Sparck Jones, Karen. Automatic Indexing. *Journal of Documentation*: 30(4), 393-432, 1974.

Turtle, H. Natural language vs. Boolean query evaluation: A comparison of retrieval performance. In W. Bruce Croft & C. J. van Rijsbergen ed. *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval*. Springer-Verlag: London, 212-220, 1994.