Journal of ICT, 3 (2), pp: 107-117

SEARCHING MALAY TEXT USING STEMMING ALGORITHM

**Rizauddin Saian and **K. R. Ku-Mahamud

*Department of Mathematical Sciences and Statistics,
Universiti Teknologi MARA Perlis, 02600 Arau, Perlis.

**Faculty of Information Technology,
Universiti Utara Malaysia, 06010 Sintok, Kedah.

Email: *rizauddin@perlis.uitm.edu.my, ** ruhana@uum.edu.my*

ABSTRACT

Stemming is an important process to improve performance of a search engine by reducing the variant word forms to common forms. This paper evaluates the retrieval effectiveness of stemming algorithm in searching and retrieving relevant Malay web pages based on user natural query words. The retrieved web pages are weighted and ranked using inverse document frequency function. The retrieval effectiveness is measured using standard recall and precision. Experiments performed show that searching with stemming improves retrieval effectiveness when compared to searching without stemming algorithm.

Keywords: stemming, information retrieval, retrieval effectiveness.

1.0 INTRODUCTION

Information retrieval (IR) system such as search engines is specifically designed for users who are not familiar with the collection, the representation of the documents, and the use of Boolean operators. These systems should be able to automatically index every word in a document and rank the document. In the first case, users should be able to enter any

natural language word(s), phrase(s) or sentence(s) to the system, without the need to enter operators. This usually implies a full text IR system whereby the system should be able to rank the retrieved documents by their estimated degree or probability of usefulness to the user.

Most return documents from user queries are irrelevant to the user's need (Wen et al., 2001). Users may pose the same query to an IR system and give different relevance judgements on the retrieved documents. Word morphology is also an issue in retrieving document (Lennon et al., 1981). In this case, a request for a Malay word such as "kegilaan" (madness), will not return any result, if all relevant documents contain the word "gila" (mad).

Conflation method could be used to overcome word morphology problems. It is a computational technique developed to transform both user's search and database words into a single canonical form (Lennon et al., 1981). This method has been applied in many important areas of practical application such as data processing, IR, text editing, word processing, linguistic analysis and even in the area of molecular biology – genetic sequence analysis (Stephan, 1994).

Language independent and language dependent are two types of conflation method. Language independent conflation method are string similarity of n-gram (Freud, 1982), dynamic programming (Stephen, 1994), while language dependent conflation includes stemming (Porter, 1980) and thesaurus. String similarity of n-gram and the dynamic programming could be used to overcome the word variants or misspellings problem. String similarity of n-gram will look at the similarity of two words. An n-gram is a set of n consecutive characters extracted from a word. Similar words such as Muhamad and Mohamad will have a high proportion of n-grams in common. The dynamic programming will look at the "edit distance" of the two strings. "Edit distance" is the minimum number of point mutations required to change string 1 into string 2 like Muhamad and Mohamad. A point mutation could be change a letter, insert a letter or delete a letter.

Stemming algorithm is a successful method to conflate the morphological variants where words with the same stem are reduced to a common form (Popovic & Willet, 1992). It has been applied in numerous areas such as natural language understanding (Cercone, 1978), literary analysis (Raben & Lieberman, 1976) and IR systems (Lovins, 1968; Lennon et al., 1981). Development of stemming algorithm for free-text retrieval purposes has been performed by

Frakes (1984), Frakes (1992), Hafer & Weiss (1974), Harman (1991), Lennon et al. (1981), Lovins (1968), Niedermair et al. (1985), Porter (1980), Ulmschneider & Doszkocs (1983) and Walker & Jones (1987).

Stemming algorithm could follow all the morphological rules, which remove the longest matching suffix once or interactively and specification of detailed context-sensitive rules in order to avoid significant error rate (Lennon et al., 1981; van Rijsbergen, 1979; Popovic & Willet, 1992; Savoy, 1993). This is as simple as removing plurals, past and present particles. According to Porter (1980), the removal of suffixes is found to be sufficient for the purpose of IR.

Besides English, stemming algorithm had also been applied to other language such as Turkish (Ekmekçioglu et al., 1996), Malay (Zainab & Nurazzah, 2004; Ahmad et al., 1996; Idris & Syed Mustapha, 2001). Zainab and Nurazzah (2004) have found that the combination of stemming and thesaurus produced the best result in retrieving documents when compared to exact match. Furthermore, various stemming algorithms for European languages have been proposed (Frakes, 1992; Kantrowitz et al., 2000; Paice, 1990; Popovic & Willet, 1992; Porter, 1980; Savoy, 1993). The designs of these stemmers range from the simplest technique, such as removing suffixes by using a list of frequent suffixes, to a more complicated design which uses the morphological structure of words in the inference process to derive a stem. The effectiveness of stemming in an IR systems depends on the morphological complexity of the language (Pirkola, 2001).

The first stemming algorithm for the Malay words was developed by Othman in 1993 (Asim, 1993). The algorithm used 121 rules which defined prefixes, suffixes, infixes and prefix-suffix pairs. This algorithm adopted a rule-based approach that slowed the stemming process. The stemmer also did not take into consideration that the stemmers would remove affixes from the root words, because the dictionary is not checked until after the first rule has been applied to the word. For example, the word "tempatan" would be stemmed to "tempat" where "an" is considered as a suffix. These two words have different meaning and may affect the performance of the IR system.

Ahmad et al. (1996) enhanced the work of Asim (1993) and come out with a set of morphological rules as the basis for the development of two new rule sets. The first set contained 432 rules of affixes and the second set contained 561 rules of affixes. The input word is first checked against the dictionary which could reduce the risk of over stemming. Experiments were done to the Quran

and research abstract data sets and results obtained showed a significant improvement over previous work.

Idris & Syed Mustapha (2001) modified the algorithm of Ahmad et al. (1996) by reducing the number of rules. They implement only the most important rules from the two patterns of the rules which are prefix and suffix rules. Infix and prefix-suffix were omitted. A new rule which they named Rule 2 was introduced in order to reduce the problem of spelling variations and exceptions in the root word. They have also found that if the prefix tests are done before the suffix test, over stemmed problem could be reduced.

In this study, an IR system was developed to retrieve Malay words from a website which contains law documents. Section 2 of this paper describes the stemming process while Section 3 details the retrieving process. Experiments performed and the performance measurements of the retrieval process are presented in Section 4 and concluding remarks are given in Section 5.

2.0 THE STEMMING PROCESS

Stemming is an important process in the retrieval system whereby a base maybe extended by one or more affixes. Affixes may be classified as prefixes, suffixes and circumfixes. The morphological description given here is taken mainly from "Tata Bahasa" (1998).

The most common prefixes in Malay include me_N, pe_N, di, ter, ber, per, se, sese, juru, ke, bel, dwi, mono, pel, pra, pro and sub. This list of prefixes with the exception of me_N and pe_N do not result in any morphographemic changes. Prefixes me_N and pe_N take on different forms (its allomorphs) depending on the initial segment of the root forms. For example, for the prefix me_N, its allomorphs are shown in Table 1. The same rules hold for the prefix pe_N and its allomorphs.

Suffixes are merely attached to the root form without any changes being made to the suffix nor the root form. Four layers of suffixes are possible in Malay as listed in Table 2. Circumfixes are discontinuous combinations of prefixes and suffixes. Listed in Table 3 below are the most common circumfixes in Malay.

Table 1: Prefix me_N and its allormorphs

me	used with the letters l, m, n, ng, ny, r, w, y
mem	used with the letters b, p, f, v (f and p dropped)
men	used with the letters d, t, c, j, z (t dropped)
meng	Used with the letters vowels, g, h, k (k dropped)
meny	used with the letter s (s dropped)
menge	for monosyllabic forms such as cat (mengecat)

Table 2: For layers of suffixes

an, wan, wait, man, is, isme, wi, in, ilah, iah, at
i, kan
mu, ku, kau, nya
lah, kah

Table 3: Circumfixes

ke_an	pe_an	se_an	men_kan
men_i	memper_i	memper_kan	per_an
se_nya	ber_an	ber_kan	di_i
di_kan			

3.0 THE RETRIEVAL SYSTEM

The retrieval system that has been developed consists of a Malay website, a database that stores the retrieved documents and a retrieval engine. The system was written using PHP 4 language and Apache version 2 was used as the web server. MySQL 4 was used to store the data. A new collection of law documents (Zeti, 2004) is set up since there is no document collection in Malay language available in standard form such as Text Retrieval Conferences (TREC), which is designed by the United States National Institute of Standards and Technology. This collection of document was then converted into a website of 86 web pages and will be used to provide the document to be retrieved in this study.

The retrieval process begins with a set of information requests (queries) where each query is a description of the information that is needed and is constructed in natural language. Stemming is applied to the queries so that the correct word is used to retrieve the document.

Document indexing is performed on the retrieved documents. Document indexing process consists of the stemming process, vector representation, weight assignment and word or term ranking. Words such as "yang", "dan" and "ialah" and hyphen are removed. Words and terms are sometimes converted from upper-case to lower-case characters and vice versa.

The retrieved document and query are represented in vector form which consist of the weights of all the word/term in a particular document or query. Vector for the ith document (D_i) and query (Q) are given as

$$D_i = \left(w_{d_{i1}}, w_{d_{i2}}, \ldots, w_{d_{it}}\right)$$
$$Q = \left(w_{q1}, w_{q2}, \ldots, w_{qt}\right)$$

where $w_{d_{i1}}$ represents the weight for the first term/word in the ith document and w_{q1} represents the weight for the first term/word in the query. For example, if the first document contains five words and the weights of the words are 0.1, 0.9, 0.3, 0.2; then $D_1 = (0.2, 0.1, 0.9, 0.3, 0.2)$. The weight is calculated as given below.

$$w_{ik} = tf_{ik} \times idf_k$$

tf_{ik} is the frequency of kth term in document D_i and idf_k is $\log\left(\frac{N}{n_k}\right)$. N is the total number of documents in collection C and n_k is the number of documents in collection C that contains the kth term. The weights are then normalized so that longer documents are not unfairly given more weight. Weight normalization is calculated as follows.

$$w_{ik} = \frac{tf_{ik} \times idf_k}{\sqrt{\sum_{k=1}^{t} (tf_{ik})^2 \left[\log\left(\frac{N}{n_k}\right)\right]^2}}$$

If a term k is absent in a document i, the weight w_{ik} will be given the value zero.

The similarity ($sim\ (Q, D_i)$) between the query and the retrieve document is calculated in order to rank the retrieved document. Similarity is calculated as follows:

$$sim\ (Q, D_i) = \frac{\sum_{k=1}^{t} w_{qj} \times w_{d_{ij}}}{\sum_{j=1}^{t} (w_{qj})^2 \sum_{j=1}^{t} (w_{d_{ij}})^2}$$

The document with the highest similarity value is considered the most related to the query. The URL of this document will be placed at the top of the return result.

4.0 FINDINGS

The traditional Average Precision-Recall measure (van Rijsbergen, 1979) is used to measure the performance of the retrieval system. The measurements are considered standard measurement by Frakes (1992). Precision (P) is the proportion of relevant items/words retrieved, while recall (R) is the proportion of retrieval items that are relevant. Precision and recall are calculated as follows.

$$P = \frac{\text{the number of retreived relevant web pages}}{\text{the number of retreived web pages}}$$

$$R = \frac{\text{the number of retreived relevant web pages}}{\text{the number of relevant web pages}}$$

The P-R measurement is based on the average precision at certain recall levels. By assuming that a certain recall level must be attained for every query, the best retrieval system is the one that attains this recall level with the fewest number of non-relevant documents. This would indicate the highest precision.

Retrieval effectiveness (E) is then calculated to reflect the relative importance of recall and precision. It is a weighted combination of recall and precision and is given by:

$$E = 1 - \frac{(1+\beta)^2 PR}{\beta^2 P + R}$$

where β is the trade off value between precision and recall.

Table 4 depicts the results of the experiments performed where stemming process is used as compared to when stemming is not incorporated in the retrieval process. Testing was done using the set of queries and relevant documents that have been constructed. Eleven queries were performs for both approaches and results from the queries were compared to the set of the actual relevant documents.

Table 4: Recall, Precision and Retrieval Effectiveness

Query	Recall		Pecision		Retrieval Effectiveness %	
	Without Stemming	Stemming	Without Stemming	Stemming	Without Stemming	Stemming
1	1.00	1.00	0.90	0.90	2.17	2.17
2	0.67	0.83	0.40	0.50	41.18	26.47
3	0.00	1.00	0.00	0.10	0.00	64.29
4	0.50	0.75	0.20	0.30	61.54	42.31
5	0.78	0.78	0.70	0.70	23.91	23.91
6	0.88	0.88	0.70	0.70	16.67	16.67
7	1.00	1.14	0.70	0.80	7.89	5.26
8	0.82	0.09	0.90	1.00	16.67	88.89
9	0.60	1.00	0.30	0.50	50.00	16.67
10	1.00	1.00	0.90	0.90	2.17	2.17
11	0.88	1.00	0.70	0.80	16.67	4.76
12	0.63	0.88	0.50	0.70	40.48	16.67
Avg	**0.73**	**0.86**	**0.58**	**0.66**	**23.28**	**25.85**

The average of the retrieval effectiveness increased by 2.57% through the used of stemming in retrieving documents. The average values for precision (66%) and retrieval (86%) are still low due to the small amount of the collected document. The values will get higher if a bigger size website is used. From the experiment, it is clear that the used of stemming algorithm in this retrieval process from a Malay website proved to be effective.

5.0 CONCLUSION

An IR system to retrieve Malay words from a website which consists of law documents has been developed. The system incorporates stemming algorithm to overcome the morphological variants Experiments in the form of queries were performed and it is observed that the performance of the retrieval process which includes recall, precision and retrieval effectiveness is better when compared to a retrieval system that is not based on stemming algorithm.

Future work could be done on IR from bigger Malay websites such as Utusan Malaysia and Berita Harian. Other aspects of retrieval performances such as database coverage, query response time and user efforts can also be include in future.

REFERENCES

Ahmad, F., Yusoff, M. & Sembok, T.M.T. (1996). Experiments with a Stemming Algorithm for Malay Words. *Journal of the American Society for Information Science, 47(12),* 909-918.

Asim O. (1993). Pengakar Perkataan Melayu Untuk Sistem Capaian Documen. MSc Thesis. National Universiti of Malaysia.

Cercone, N. (1978). Morphological Analysis and Lexicon Design for Natural Language Processing. *Computers and Humanities, 11,* 199-209.

Ekmekçioglu, F. Çuna, L, Michael F. & Willett, P. (1996). Stemming and N-gram matching for term conflation in Turkish texts. Information Research, 1(1). Retrieved July 20, 2004 from http://informationr.net/ir/2-2/paper13.html.

Frakes, W.B. (1992). Stemming Algorithms. In W. B. Frakes and R. Baeza (Ed.), *Information Retrieval, Data Structures and Algorithms.* Upper Saddle River, NJ, USA: Prentice Hall.

Frakes, W.B. (1984). Term Conflation for Information Retrieval. In van Rijsbergen, C. J. (Ed.), *Research and Development in Information Retrieval.* CUP: Cambridge.

Freud, G.E. & Willett, P. (1982). Online Identification of Word Variants and Arbitrary Truncation Searching Using a String Similarity Measure. *Information Technology Research and Development, 1*, 177-187.

Hafer, M.A. & Weiss, S.F. (1974). Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval, 10*, 371-385.

Harman, D. (1991). How Effective is Suffixing?. *Journal of the American Society for Information Science, 42(1)*, 7-15.

Idris, N. & Syed Mustapha, S.M.F.D. (2001). Stemming for Term Conflation in Malay Texts. *International Conference of Artificial Intelligence.* April 23, 2001. (pp. 1512 – 1517). Las Vegas.

Kantrowitz, M., Mohit, B., & Mittal, V. (2000). Stemming and Its Effects of TFIDF Ranking. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* July 24-28, 2000. (pp. 357-359). Athens, Greece.

Lennon, M., Peirce, D.S., Tarry, B.D. & Willet, P. (1981). An evaluation for some conflation algorithms for information retrieval. *Journal of Information Science, 3*, 177-183.

Lovins, J.B. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics, 11*, 22-31.

Niedermair, G.T., Thurmair, G. & Buttel, I. (1984). MARS A Retrieval Tool on the Basis of Morphological Analysis. In van Rijsbergen, C. J. (Ed.), *Research and Development in Information Retrieval, Proceedings of the Third Joint BCS/ACM Symposium on Research and Development in Information Retrieval, Cambridge.* July 2-6, 1984. (pp. 369-381). CUP: Cambridge, England.

Paice, C.D. (1990). Another Stemmer. *ACM SIGIR Forum, 24(3)*, 56-61.

Pirkola, A. (2001). Morphological Typology of Languages for IR. *Journal of Documentation, 57*, 330-348.

Popovic, M. & Willet, P. (1992). The Effectiveness of Stemming for Natural-Language Access to Slovene Textual Data. *Journal of the American Society for Information Science, 43*(5), 384-390.

Porter, M.F. (1980). An Algorithm for Suffix Stripping. *Program*, *14*(3), 130-137.

Raben, J. & Lieberman, D.V. (1976). Text comparison: principles and a program. In Jones, A & Churchouse, R. F. (Eds.), *The computer in literacy and linguistic studies* (pp. 297-308). Cardiff: University of Wales Press.

Savoy, J. (1993). Stemming of French Words Based on Grammatical Categories. *Journal of the American Society for Information Science*, *44*, 1-9.

Stephen, G.A. (1994). String Searching Algorithm. In *Lecture Notes Series on Computing*. Singapore: World Scientific Publishing Co. Pte. Ltd.

Ulmschneider, J.E. & Doszkocs, T. (1983). A Practical Stemming Algorithm for Online Search Assistance. *Online Review*, *7*, 301-318.

Van Rijsbergen, C.J. (1979). *Information Retrieval* (Second Edition). London: Butterworths.

Walker, S. & Jones, R.M. (1987). Improving Subject Retrieval in Online Catalogues. *1. Stemming, Automatic Spelling Correction and Cross-Reference Tables*, British Library Research Paper, London.

Wen, J.R., Nie, J.Y. & Zhang H.J. (2001). Clustering User Queries of a Search Engines. *ACM, Proceedings Of The Tenth International Conference On World Wide Web*. May 1-5, 2001. (pp. 162-168). Hong Kong.

Zainab, A.B. & Nurazzah, A.R. (2004). Evaluating the Effectiveness of Conflation Methods in Retreiving Malay Translated Al-Quran Texts and Images. *Conference on Scientific and Social Research, UiTM*. May 19-21, 2004. Kuching, Sawarak, Malaysia.

Zeti, Z.M.Z. (2004). Penipuan Kad Kredit di Malaysia. LLM Thesis, National University of Malaysia.