

Word Stemming Algorithms and Retrieval Effectiveness in Malay and Arabic Documents Retrieval Systems

Tengku Mohd T. Sembok

Abstract—Documents retrieval in Information Retrieval Systems (IRS) is generally about understanding of information in the documents concern. The more the system able to understand the contents of documents the more effective will be the retrieval outcomes. But understanding of the contents is a very complex task. Conventional IRS apply algorithms that can only approximate the meaning of document contents through keywords approach using vector space model. Keywords may be unstemmed or stemmed. When keywords are stemmed and conflated in retrieving process, we are a step forwards in applying semantic technology in IRS. Word stemming is a process in morphological analysis under natural language processing, before syntactic and semantic analysis. We have developed algorithms for Malay and Arabic and incorporated stemming in our experimental systems in order to measure retrieval effectiveness. The results have shown that the retrieval effectiveness has increased when stemming is used in the systems.

Keywords—Information Retrieval, Natural Language Processing, Artificial Intelligence.

I. INTRODUCTION

INFORMATION Retrieval (IR) can be defined broadly as the study of how to determine and retrieve from a corpus of stored information the portions which are relevant to particular information needs. Let us assume that there is a store consisting of a large collection of information on some particular topics, or combination of various topics. The information may be stored in a highly structured form or in an unstructured form, depending upon its application. A user of the store, at times, seeks certain information which he may not know to solve a *problem*. He therefore has to express his *information need* as a request for information in one form or another. Thus IR is concerned with the determining and retrieving of information that is relevant to his information need as expressed by his *request* and translated into a *query* which conforms to a specific information retrieval system(IRS) used. An IRS normally stores *surrogates* of the actually *documents* in the system to represent the documents and the *information* stored in them [1].

Manuscript received November 20, 2005. This work was supported in part by the Malaysian IRPA Grant 04-02-02-027.

T.M.T.Sembok is with the National University of Malaysia, Bangi 43600, Selangor, Malaysia (phone: +60123379476; fax: +60389216732; e-mail: tmts@pkrisc.ukm.my).

II. HUMAN INFORMATION-PROCESSING MODEL AND IRS MODEL

When a person reads documents to seek for information which are relevant to his needs to solve a problem, he is engaging himself in a highly intellectual process: reading documents written in natural language, using his working memory, and accessing his long term memory in order to understand the documents and decide which are relevant and which are not. This cognitive process of determining the degree of relevance of documents can be expressed based on human information-processing model [2] as depicted in Fig. 1.

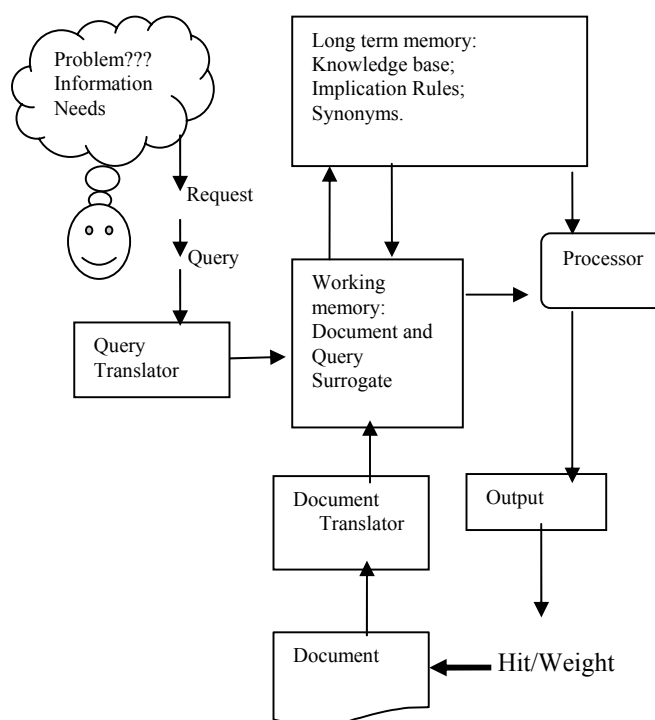


Fig. 1 Information Retrieval System Model

III. SURROGATES AND REPRESENTATION

In conventional document retrieval systems, the surrogates of documents and queries are built by an unstructured collection of simple descriptors, i.e. the keywords.

In conventional document retrieval systems, documents are represented by sets of keywords, or index terms of the form

$$D_i = (t_1, w_{i1}; t_2, w_{i2}; \dots; t_n, w_{in})$$

where w_{ij} represents the value or weight of term t_j which is assigned to document D_i . In the Boolean model, the terms are unweighted. Thus, the values of the w_{ij} are restricted to either 0 or 1 for terms that are respectively absent from, or present in, a given document. The vector space model gives certain value to w_{ij} which reflect the importance of the term in the document concern. The basic weights for the terms throughout the document collection are normally calculated using statistical techniques.

In Boolean model, the requests are expressed as Boolean combinations of index terms using logical operators and, or, and not. For example, a query Q might be expressed as $Q = ((t_i \text{ and } t_j) \text{ or } t_k)$.

In response to the query given above, all documents indexed either by the combination of t_i and t_j , or by t_k would be retrieved.

In vector space model query is represent by the vector:
 $Q = (t_1, q_1; t_2, q_2; \dots; t_n, q_n)$

where w_{ij} reflects the present and the importance of the term t_i in the query. There are many ways of computing similarity coefficients between a given query and each stored document. For example, one can use the well-known inner-product function, as follows, to do the matching:

$$\text{Inner Product Function: } \text{similarity}(D_i, Q) = \sum_{j=1, n} (q_j \cdot w_{ij})$$

In these models we need techniques to code and to match strings of keywords that represent the documents and the query.

IV. CONFLATION METHODS

Common to all languages, words variants are formed by the usage of affixes, alternative spelling (new and old spellings), multi-word concepts, transliteration, abbreviations and spelling errors [3]. Examples: compute computes, computed, computer, computers, computerise, computation, computational, etc.

These problems can be solved by the development of computational technique that could transform both user's search and database words into a single canonical form [4] that is known as Conflation. Conflation is an important facility in any text retrieval system, whereby, it is also able to find not identical words in the database that match the vocabulary that the user used [5].

Stemming algorithm is used to conflate morphological variants. For English we Porter's Algorithms [6] which stems only suffixes. For Malay we have Othman's Algorithm [7] 1993) which has 121 rules for prefixes, suffixes and infixes; Fatimah's Algorithm [8] which has 561 rules and a root words dictionary; and Norsiah's Algorithm which uses neural network.

Problems faced in stemming of Malay words [8][9] Sembok et al, and Fatimah et al. are categorized as

understemming, overstemming, unchanged and spelling exception as in the examples given below:

Understemming:

kedudukan => keduduk (duduk)

Overstemming:

pemakanan => mak (makan)

kesan => kes (kesan)

Unchanged:

masalah => masalah (masa)

gerigi => gerigi (gigi)

Spelling exception:

mengandungi => gandung (kandung)

For Arabic we have Hani's Algorithm[10] and Belal's Algorithm [11] with rules which are more complex especially in dealing with infixes.

V. EXPERIMENTS

An experiment to compare retrieval effectiveness of using conflation method, stemming of Malay words, has been carried out and found that it performs better than non-conflation method. The graph in Figure 2 show the analysis of performance based on recall-precision measurement between the conflation methods, Rule order application and Longest Match, against non-conflation method on a collection of Malay texts.

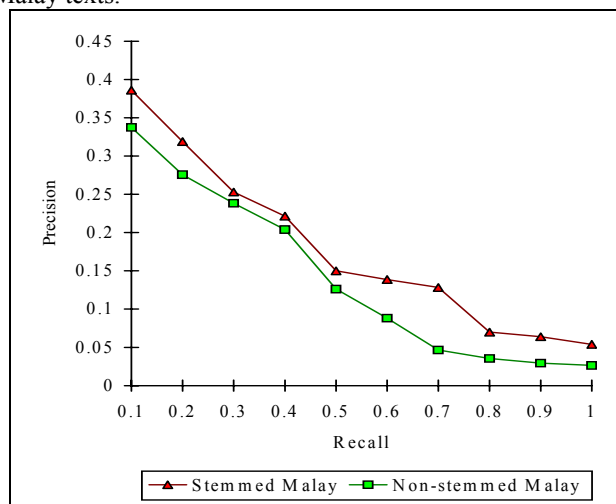


Fig. 2 Average Recall-Precision Graph for conflation and non-conflation methods on Malay Texts

A similar experiment on a collection of Arabic texts has been carried and found a similar result as depicted in Fig. 3 below.

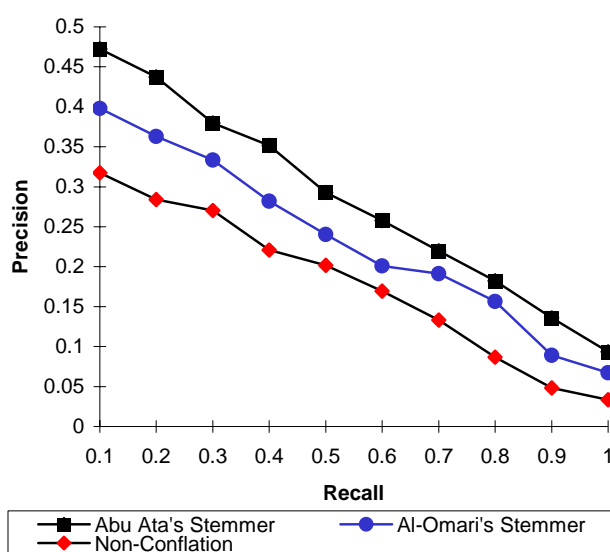


Fig. 3 Average Recall-Precision Graph for conflation and non-conflation methods on Arabic Texts

VI. SYSTEMS DEVELOPED

A few multilingual systems have been developed using the techniques and the tools investigated, among which are:

1. Malay-English/English-Malay Scientific Terms Database (Stand-alone and Web based) - used searching technique based on combination of stemming and n-grams [12].
2. Al-Faruq - Malay-English-Arabic Al-Quran Database [10].
3. SCAQ - Malay/English Al-Quran on the Web: Malay and English queries can be fused to enhance retrieval of information [13].
4. SISDOM'98 - Malay/English Documents Retrieval System: using stemming methods and tf-idf weighting [14].

Analysis on the performance of the system has been carried out to study the behavior of the system as regard to the usage of different languages. One of the findings obtained, shows that the same query written in different languages do not retrieve the same hits list. Only around 50% of the documents are overlap.

VII. FURTHER WORK

With stemming algorithms in place we can further with syntactic and semantic analysis of document to enhance retrieval effectiveness. This shall be our next focus of research.

REFERENCES

- [1] Mizzaro, S. Relevance: The Whole History. *Journal of American Society of Information Science*, Vol.48, No.9, 1997. pp.810-832.
- [2] Gagne, E.D., Yekovich, C.W., Yekovich, F.R. *The Cognitive Psychology of The School Learning*, Harper Collin. 1993.

- [3] Freund, G.E. & Willett, P. Online identification of word variants and arbitrary truncation searching using a string similarity measure. *Information Technology: Research and Development 1*: 1982. 177-187.
- [4] Lennon, M., Pierce, D., Tarry, B. & Willett, P. An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science 3*: 1981. 177-183.
- [5] Ekmekcioglu, F.C., Lynch, M.F., Robertson, A.M., Sembok, T.M.T. & Willett, P. Comparison of n-gram matching and stemming for term conflation in English, Malay, and Turkish texts. *Text Technology: The Journal of Computer Text Processing 6*(1): 1996. 1-14.
- [6] Porter M.F. An Algorithm for suffix stripping, *Program*, 14(3), 1980. pp.130-137.
- [7] Othman, A. Pengakar perkataan melayu untuk sistem capaian dokumen. MSc Thesis. National University of Malaysia. 1993.
- [8] Fatimah Ahmad, Mohammed Yusoff, Tengku Mohd. T. Sembok. "Experiments with A Malay Stemming Algorithm", *Journal of American Society of Information Science*. 1996.
- [9] Sembok, T.M.T, Yusoff, M. & Ahmad, F. A malay stemming algorithm for information retrieval. *Proceedings of the 4th International Conference and Exhibition on Multi-lingual Computing*. 1994. 5.1.2.1-5.1.2.10.
- [10] Hani Moh'd Al-Omari, Tengku Mohd. T. Sembok, Mohammed Yusoff, ALMAS: An Arabic Language Morphological Analyser System, *Malaysian Journal of Computer Science*, Vol. 8, no.2, University of Malaya. 1995.
- [11] Belal Abu Ata, Tengku Mohd T. Sembok, Mohamed Yusoff. *Implementations of A Malay Stemming Algorithm Using Hashing Technique*, *Proceedings of the ICIMU'98: International Conference on Information Technology and Multimedia*, UNITEN, 28-30 Sept. 1998.
- [12] Sembok, Tengku Mohd Tengku. Application of Mathematical Functional Decomposition in Document Indexing, *Prosiding : Pengintegrasian Teknologi dalam Sains Matematik*. Penang: USM. 1999.
- [13] Saidah Saad. 1998. Pembangunan dan Eksperimen ke atas satu sistem capaian maklumat Al-Quran dwi bahasa berasaskan Web. MSc. Thesis. UKM.
- [14] Sembok, T.M.T. & Willett, P. Experiments with n-gram string-similarity measure on malay texts. *Technical Report. Universiti Kebangsaan Malaysia*. 1995.