Effectiveness of Stemming and n-grams String Similarity Matching on Malay Documents

Tengku Mohd T. Sembok and Zainab Abu Bakar

Abstract— There are two main classes of conflation algorithms, namely, string-similarity algorithms and stemming algorithms. String-similarity matching algorithms, bi-grams and tri-grams, are used in the experiments conducted on Malay texts. Malay stemming algorithms used in the experiments is developed by Fatimah et al. Inherent characteristics of n-grams on Malay documents are discussed in this paper. Retrieval effectiveness experiments using several variations of combinations between n-grams and stemming algorithms are performed in order to find the best combination. The variations experimented are: both nonstemmed queries and documents; stemmed queries and nonstemmed documents; and both stemmed queries and documents. Further experiments are then carried out by removing the most frequently occurring n-grams. Besides using dice coefficients to rank documents, inverse document frequency (*idf*) weights are also used. Interpolation technique and standard recall-precision functions are used to calculate recall-precision values. It is found that using combined search, n-gram matching and stemming, improves retrieval effectiveness. Removing the most frequently occurring n-gram that appears in about 46% of the words also improve the retrieval effectiveness.

***Keywords*—**Information retrieval, string similarity matching, stemming algorithms.

I. Introduction

Information are increasingly available for user searching either through online connections to remote hosts or locally on CD-ROM. There is a wide range of text-based information which are usually textual and may also contain other types of information such as photographs, graphs, voices and animation that can be searched and retrieved especially through the internet. This has led to an increased number of new researchers into the area of information retrieval.

The study on information retrieval is focused on how to determine and retrieve from a corpus of stored information, the portion which is relevant to particular information needs [1]. Common to all languages, text-based information systems that use free text for indexing and retrieval, have variation in word formation. This is due to the usage of affixes, alternative spelling, multi-word concepts, transliteration, abbreviations, and spelling errors [2]. Many methods have been described to overcome certain types of word variant using conflation algorithms [3]. Conflation is defined as a computational procedure which identifies word variants and reduced them to a single canonical form [4]. Conflation algorithms are broadly classified into two main classes: stemming algorithm which is language dependent and string-similarity algorithm which are language independent.

Stemming Algorithm

A stemming algorithm is a procedure that reduce words with the same stem to a common form, usually by removing derivational and inflectional suffixes from each word [5]. For example, the words ***study***, ***studies***, ***studied***, ***studying***, ***student*** or ***studious*** are reduced to the root wood ***study***. In information retrieval, grouping words into common form will increase in retrieving relevant documents against a given query[6]. The development of stemming algorithms for free-text retrieval purpose is evidenced by the work of many researchers [7].

Most of these studies have focused on the development and use of stemming algorithms for English and similar languages such as Slovene and French. Stemming algorithms can be very simple by just removing plurals, past and present particles to very complex techniques that include all morphological rules. Such complex procedures require either removal of the longest matching suffix once or interatively and specification of detailed context-sensitive rules in order to avoid significant error rate [7][8][9].

String-Similarity Matching

The removal of suffixes by a stemmer in English and similar languages, Slovene and French, are found to be sufficient for the purpose of information retrieval [10] but this not so in Malay [11] and Arabic [12]. To stem Malay text effectively, not only suffixes but prefixes and infixes must be removed in proper order as described by Ahmad et al.[13].

The stemming procedures in English and similar languages are generally unsuited to the conflation of all possible types of word variant and they show specific defects in chosen applications [2][13]. A very different approach to the problem of variant known as n-gram similarity measures was devised by Adamson and Boreham [14]. This approach suggested that words which have a high degree of structural similarity tend to be similar in meaning. Each word is represented by a list of its constituent n-grams, where *n* is the number of adjacent characters in the substrings. Using these lists, similarity measures between pair of words are calculated based on shared unique n-grams and the number of unique n-grams for each word. Typical values for *n* are 2 or 3, which correspond to the use of bigrams and trigrams. For bigram the number of n-grams is *n*+1, and trigram is *n*+2. A quantitative similarity measure *S* between them can

Manuscript received January 23, 2011: Revised version received April 2, 2011.

T.M.T. Sembok is with Universiti Kebangsaan Malaysia, Bangi, 43600 Malaysia (phone:+60123373539; fax: +60389253276; e-mail: tmts@ftsm.ukm.my).

Z.A. Bakar is with Universiti Teknologi MARA, Shah Alam, Malaysia (e-mail: zainab@fsmsk.uitm.edu.my).

be computed by using Dice Coefficient or the Overlap Coefficient as shown in Table-1 below.

Table-1: Similarity Measure of Bigram and Trigram for words **construct** and **destruct**.

	Bigram	Trigram
Unique N-gram of Word 1	***c co on ns st tr ru uc ct t***	****c *co con ons nst str tru ruc uct ct* t****
Unique N-gram of Word 2	***d de es st tr tu uc ct t***	****d *de des est str tru ruc uct ct* t****
A = Unique N-gram of Word 1	10	11
B = Unique N-gram of Word 2	9	10
C = Shared Unique Bigram	6	6
Dice Coefficient= (2C)/(A+B)	0.631579	0.571429
Overlap Coefficient = C/min(A,B)	0.666667	0.600000

Note : * denotes a space

Experiments on N-gram Matching

Again there are much research on n-gram string-similarity measures are done on English. Adamson and Boreham [14] used inter similarity coefficient to cluster a small group of mathematical titles on the basis of their constituent digrams and form the basis for a document retrieval system; but the same procedure gave poor results using Cransfield test collection [15]. Using the same collection Lenon *et al* [4] represented all unique terms occuring in the document titles and queries by lists of constituent digrams and trigrams. Index terms with a similarity coefficient greater than some threshold were considered to be variants of the query term and are used for searching relevant documents.

Freund and Willett [2] performed online query expansion using inverted file structure constituted of digrams on Evans and Vasmani test collections. They found that using the digram inverted file retrieved more non-related terms at lower similarity threshold compared to terms retrieved using the trigram inverted file. They then used arbitrary truncation of the query terms and retrieved higher proportion of related words and still maintain an acceptable level of precision.

Robertson and Willett [16] used n-gram matching, both digram and trigram, phonetic and non-phonetic coding, and dynamic programming methods to identify words in the Historical Text Databasea using query words which are modern English. They concluded that digram string-similarity is the appropriate method to be implemented in an operational environment where a large dictionary was to be serached.

Other sudies on different languages include Turkish and Malay. Ekmekcioglu *et al.*[17] had performed n-gram string-similarity matching experiments similar to that of Lenon *et al* [4] on six Turkish databases and found that trigrams performed slightly better than bigrams in most of the text corpora. They performed experiments by using stemmed queries and nonstemmed dictionary and both stemmed queries and dictionary. Results show that the stemmed versions performed significantly better than the nonstemmed.

Sembok et al. [18] performed experiment using n-gram string-similarity measure, digram and trigram on Malay dictionary and queries and found that overlap coefficient results are better than dice coefficients and that digram perform significantly better than the trigrams. They too found that stemming both queries and the dictionary performs significantly better than just stemming only the queries. However, no conclusion is made between the conventional stemmed-Boolean approach and incorporating-stemming in n-grams approach since both approaches are based on different paradigms.

Other applications using n-gram include work done by Yannakoudakis and Angelidakis [19]. They used n-grams differently that is to show distribution of n-grams in the Shorter Oxford English Dictionary for values of *n* from 2 to 5, from bigram to pentragram. They shown that the corresponding redundancy of n-grams increases from 0.1067 to 0.3409.

Experimental Detail

Objectives

The objectives of the experiments are to investigate the inherent characteristics of n-gram in Malay text and to evaluate in terms of retrieval effectiveness of n-gram matching techniques, bigram and trigram, applied to Malay text.

Experimental Setup

The Malay test collection used in the experiments is the collection of Ahmad et al. [18] which is based on Malay translation of the Quran. There are 6236 documents and 36 natural language queries.

Two types of index term dictionary (hereafter abbreviated to ITD) are created; index nonstemmed term dictionary (hereafter abbreviated to INTD) and index stemmed term dictionary (hereafter abbreviated to ISTD). Stop words are removed before the terms are indexed. From INTD, the number of unique index nonstemmed terms is 5525, and from ISTD, the number of unique index stemmed terms is 2101.

Characteristics of N-gram on Malay Text

There are 728 distinct bigrams and 18980 distinct trigrams both include leading and trailing space(s). The sizes in bytes for each of the bigram and trigram files are shown in Table 2:

Table 2. File sizes of Bigrams and Trigrams

	Bigram	Trigram
INTD	517358 bytes	858580 bytes
ISTD	145278 bytes	444340 bytes

Bigram and trigram created from INTD herafter are known as nonstemmed- bigram and nonstemmed-trigram. And those created from ISTD are known as stemmed-bigram and stemmed-trigram.

Bigram

The theoritical maximum number of distinct bigrams that can be found is 26x26=676 (without spaces) and 27x27=729 (with spaces), less 1 bigram of the form (space,space) which gives a total of 728. As in English these theoretical maxima can never be reached in practise because certain bigrams such as QQ and XY [19] simply do not occur in Malay text too. In this Malay text the maximum number of non-zero nonstemmed-bigrams is 377, only 52%, and from stemmed-bigrams 355, only 49%. Table-3 contains the top 100 bigrams for both nonstemmed and stemmed bigrams. In order to get a general idea about the whole range of both bigrams, their rank-frequency distribution and zipfian distribution were plotted (Figures 1 and 2).

Trigrams

The maximum possible number of distinct trigrams is 26x26x26=17576 (without spaces) and 27x27x27=19683 (with spaces), less 702 bigrams with a space in between, less 1 trigram of the form (space,space,space) which gave a total of 19890. Appropriate storage arrays were used to hold both nonstemmed and stemmed trigrams as indicated in Table-2 above. The total maximum number of non-zero nonstemmed-trigrams is 2214 which utilised only 12% and for stemmed-trigrams is 1892 which utilised only 10%. Thus during run time the size of array allocation can greatly be reduced.

The most 50 frequently occuring trigrams are listed in Table-4 above. The rank-frequency distribution and zipfian distribution for both trigrams are presented in Figures 3 and 4 above.

Zipfian Distribution

The curves demonstrating the zipfian distributions in Figures1,2,3 and 4, of the n-gram for Malay text have a similar hyperbolic curve complying to the Zipf's law [20] which states that the frequency of words and the rank order is approximately constant. Such analysis should not be restricted to just words [1]. Luhn [21] used Zipf's law to specify upper and a lower cut-offs. The words below the upper cut-off were considered to be common and those below the lower cut-off to be rare, leaving the rest to be significant words. These cut-offs were established by trial and error by estimating in both direction from the peak of the rank-order position. Thus by removing the most frequently n-grams should not change the result of retrieval effectiveness of the documents, as carried out later.

Experimental Evaluation Procedure

The experiments performed involved the ranking and the calculation of string similarity measures of each unique terms in the ITD to a specified query term. This procedure is the same as automatic query expansion approach as set by Lennon *et al* [4]. Following are the evaluation procedures that are carried out.

Table-3. The Most Frequently Occurring Bigrams (with spaces)

	Nonstemmed		Stemmed	
Rank	Bigra	Frequency	Bigra	Frequency
1	an	2567	an	401
2	ka	1472	a	341
3	n	1391	ng	330
4	a	1356	s	242
5	ng	1136	g	209
6	er	1111	b	208
7	ya	1104	t	207
8	m	1067	h	204
9	ny	1062	k	200
10	me	917	i	198
11	en	909	ah	195
12	ah	868	ra	191
13	la	826	la	190
14	d	731	ka	188
15	di	707	n	184
16	h	674	at	175
17	b	632	ar	174
18	k	594	t	174
19	pe	581	r	171
20	p	578	m	167
21	ak	563	ta	166
22	be	549	p	160
23	em	530	er	158
24	at	522	ak	156
25	ta	509	ba	152
26	ra	498	ma	150
27	in	471	k	144
28	ar	462	sa	139
29	u	461	in	135
30	ga	442	u	130
31	ba	441	al	129
32	i	430	am	128
33	ke	428	en	124
34	sa	419	un	121
35	t	417	as	120
36	s	414	pa	119
37	nn	395	l	117
38	ku	391	da	115
39	ha	385	ha	113
40	ma	382	a	109
41	mu	358	ur	106
42	al	349	ga	103
43	un	349	s	99
44	am	341	ya	97
45	pa	335	ri	95
46	ik	320	na	94
47	se	316	d	93
48	tu	310	se	89
49	na	308	el	88
50	as	299	h	87

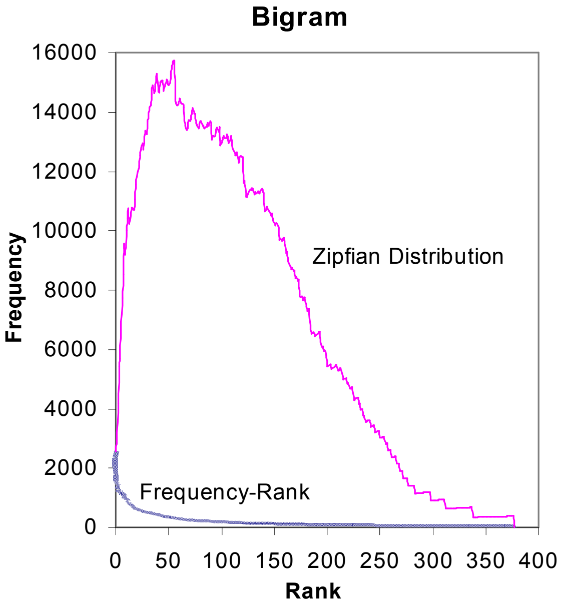

Figure 1. Nonstemmed Bigram Rank-Frequency and Zipfian Distribution

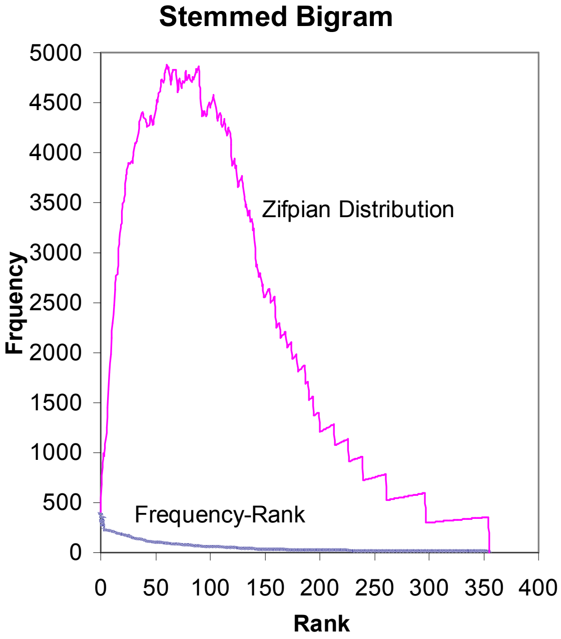

Figure 2. Stemmed Bigram Rank-Frequency and Zipfian Distribution

Table-4. The Most Frequently Occurring Trigrams (with spaces)

	Nonstemmed		Stemmed	
Rank	Trigram	Frequency	Trigram	Frequency
1	n	1391	a	341
2	a	1356	s	242
3	an	1312	g	209
4	kan	1132	b	208
5	m	1067	ng	208
6	nya	987	t	207
7	ya	942	h	204
8	me	890	k	200
9	d	731	i	198
10	h	674	n	184
11	b	632	t	174
12	di	618	r	171
13	k	594	m	167
14	p	578	ang	167
15	ah	568	p	160
16	men	531	k	144
17	ang	488	u	130
18	u	461	ah	127
19	i	430	an	120
20	be	429	l	117
21	pe	429	at	116
22	lah	418	a	109
23	t	417	s	99
24	s	414	d	93
25	ber	402	h	87
26	nny	394	se	86
27	ann	375	r	85
28	ke	367	g	81
29	eng	348	m	80
30	per	309	ar	76
31	aka	286	ba	73
32	nga	258	ke	73
33	mem	256	ma	72
34	g	241	te	71
35	ng	240	l	70
36	se	223	be	65
37	t	218	ak	65
38	mu	213	pe	60
39	emb	205	ta	60
40	ala	202	j	58
41	te	201	ka	57
42	ngk	196	sa	57
43	r	193	ung	57
44	gan	192	i	56
45	ran	183	ala	55
46	era	176	ing	54
47	ika	173	am	52
48	tan	170	n	50
49	ter	167	ha	49
50	ara	163	la	49

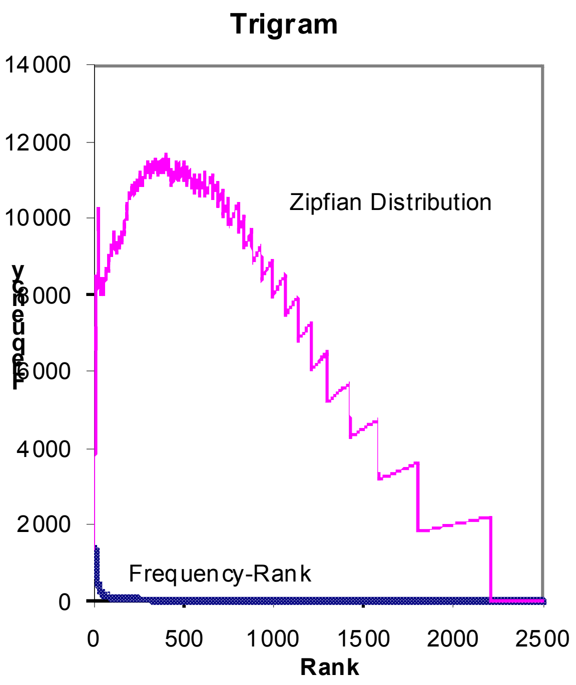

Figure 3. Nonstemmed Trigram Rank-Frequency and Zipfian Distribution

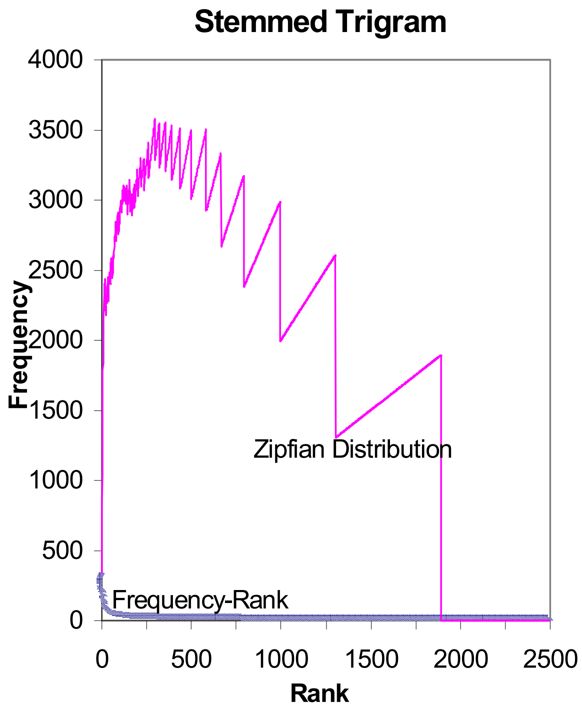

Figure 4. Stemmed Trigram Rank-Frequency and Zipfian Distribution

Threshold Value

When a query term is submitted to the system, it is broken down into its constituent n-grams. A list of index terms that have some or all common n-grams is kept. The index terms are used to calculate Dice similarity values. If the Dice similarity value of an index term is equal or greater than the preset threshold value, it will be accepted. Initially the threshold values were set to be from 0.8 down to 0.4, employing an interval of 0.05 [2]. After running a few experiments, there is not much significant changes using an interval of 0.05. The threshold values that are used in the experiments are 1.00 and 0.8 down to 0.4, employing intervals of 0.1.

Ranking Documents

The Dice coefficient is used to weight retrieved documents together with the following Inverse Document Frequency (*idf*) function (Spark Jones, 1972):

$$
\begin{aligned}
WEIGHT_{ik} &= tf \text{ x } idf \\
&= FREQ_{ik} \text{ x } \left[\log_2 \frac{n}{DOCFREQ_k} + 1 \right]
\end{aligned}
$$

where *tf* is the term frequency and *idf* is the inverse document frequency, *n* is the total number of documents in the collection, $DOCFREQ_k$ is the number of documents to which term *k* is assigned and $FREQ_{ik}$ is the frequency of term *k* in a given document *i*. For a query having more than one index term, the weight for a particular document is computed as the sum of all weights calculated to the index terms appearing in the document. The documents are then ranked in decreasing order.

Recall and Precision Values

Recall and precision values are calculated using the ranked documents and a list of relevant documents according to the query number as in the experiments carried out by Ahmad [22]. Recall, R, is defined as the proportion of relevant material retrieved , while precision, P, is the proportion of retrieved material that is relevant. The standard *R* and *P* are defined as [23]:

Precision = (Number of Documents Retrieved and Relevant)/ (Number of Documents Retrieved)

Recall = (Number of Documents Retrieved and Relevant)/ (Number of Relevant Documents)

After obtaining all the recall and precision values for each query, the precision-recall curve is calculated. This is obtained by specifying a set of standard recall values from 0.1 to 1.0, with interval of 0.1. To get unique precision value that corresponds exactly, interpolation technique (van Rjsbergen,1979) is used. This process is carried out for all 36 queries (see Appendix A). There are 36 sets of standard recall-precision values and the averages for all the queries are tabulated. These procedures are run again for the next threshold value. To complete 9 variations of the experiments total number of computer runs = 1994 (9 variations x 6 thresholds x 36 queries).

Experimental Results and Discussion

Analysis of the results obtained show that stemming both keywords and documents is clearly the best than either stemming the keywords only or do not stem the keywords at all. In fact results from both bigram and trigram search, results obtained by not applying stemming algorithm to the keywords performed better than applying stemming to the keywords.

Finally, Table-5 shows that in general applying stemming algorithm to both keywords and documents improve the average recall-precision values. There is an improvement in the retrieval effectiveness using bigram but the improvement is not significant.

Table-5. Best Average Recall-Precision Values of Various Experiments (refer the table Keys for the Experiment column)

Experiment	Threshold	Average Precision
bsksd	1.0	0.183373
tsksd	1.0	0.178227
brsksd	0.8	0.173899
tnknd	0.6	0.147814
brnknd	0.6	0.148845
tsknd	0.6	0.141466
bnknd	1.0	0.143027
bsknd	0.7	0.138249
NC		0.140768
brsknd	0.7	0.136704

Keys for Experiment: (example **bsksd** = bigram;stemmed keywords; stemmed document)

Key	Description	Key	Description
b	***b**igram*	**k**	***k**eywords*
t	***t**rigram*	**NC**	***N**on-**C**onflation*
s	***s**temmed*	**r**	***r**emoved the most frequent bigram "an"*
n	***n**onstemmed*		
d	***d**ocument*		

For the detail evaluation of the experimental result, see Table-6 in Appendix B.

Conclusions

In this paper inherent characteristics of bigram and trigram are discussed. Experiments using various combinations of bigram, trigram and stemming algorithms are performed on Malay queries and documents. Further experiments are then carried out by removing the most frequently occurring bigram. The experiments show that using combined search, *n*-gram matching and stemming, improves retrieval effectiveness. Removing the most frequently occurring *n*-gram that appears in about 46% of the words also improve the retrieval effectiveness.

Stemming and ngrams approach have been applied in the Malay-English Terminology Retrieval System by Sembok et al.[24] to retrieve the Malay science terminology. Other experiments on retrieval effectiveness on Malay texts are reported by Hamzah and Sembok [25] which experimented on various matching algorithm such as cosine, dice, and overlap similarity matching. More advance processing on Malay texts retrieval based on semantic approach and and specific domain as performed by Stanojević and Vraneš [26] and Pohorec et al. [27] shall be our next project. This undertaking shall enable us to proceed further into question answering system for Malay language [28].

References

[1] van Rijsbergen, C.J. 1979. *Information Retrieval.* 2nd edition. London: Butterworths.

[2] Freund, G.E. and 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.

[3] Hall, P.A.V. and Dowling, G.R. 1980. Approximate string matching. *Computing Surveys* **12**: 381-402.

[4] Lennon, M. , Peirce, D.S., Tamy, B.D. and Willett, P. 1981. An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science* **3**: 177-183.

[5] Lovins, J.B. 1968. Development of a stemming algorithm. *Mechaniacl Translation and Computational Linguistics* **11**:22-31.

[6] Harman, D. 1991. How effective is suffixing?. *Journal of the American Society for Information Science* **42**: 7-15.

[7] Popovic, M. and Willett, P. 1992. The effectiveness of stemming for natural language access to Slovene textual data. *Journal of the American Society for Information Science* **43**: 384-390.

[8] Lennon, M. , Peirce, D.S., Tamy, B.D. and Willett, P. 1981. An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science* **3**: 177-183.

[9] Savoy, J. 1993. Stemming of French words based on grammatical categories. *Journal of the American Society for Information Science* **44**(1) : 1-9.

[10] Porter, M.F. 1980. An algorithm for suffix stripping. *Program* **14**(3): 130-137.

[11] Sembok, T.M.T., Yusoff, M. And Ahmad, F. 1994. A Malay stemming algorithm for information retrieval. *Proceedings of the 4th International Conference and Exhibition on Multi-Lingual Computing.* London.

[12] Al-Kharashi, I.A. & Evens, M.W. 1994. Comparing words, Stems, and Roots as Index Terms in an Arabic Information Retrieval System. *Journal of the American Society for Information Science*. **45**(8): 548-560.

[13] Ahmad, F., Yusoff, M, and Sembok, T.M.T. 1995. Experiments with a Malay Stemming Algorithm. JASIS.

[14] Adamson, G.W. and Boreham, J. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval* **10** : 253-260.

[15] Willett, P. 1979. Document retrieval experiments using vocabularies of varying size. II. Hashing, truncation, digram and trigram encoding of index terms. *Journal of Documentation* **35**:296-305.

[16] Robertson, A.M. (1992). An evaluation of algorithmic techniques for the identification of word variants in historical text database. Ph.D. Thesis. Department of Information Studies, University of Sheffield.

[17] Ekmekcioglu, F.C., Lynch, M.F., and Willett, P. 1995. Language processing techniques for the implementation of a document retrieval system for Turkish text databases. *Department of Information Studies:University of Sheffield.*

[18] Sembok, T.M.T., Palasundram, K., Ali, N.M., Aidanismah, Y., Wook, T.M.T. 2003. Istilah Sains: A Malay-English Terminology Retrieval System Experiment Using Stemming and N-grms Approach on Malay Words, Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access, 6th International Conference on Asian Digital Libraries, ICADL 2003, Kuala Lumpur. Berlin: Springer.

[19] Yannakoudakis, E.J. and Angelidakis, G. 1988. An Insight into the Entropy and Redundancy of the English Dictionary. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence **10**(6): 960-970.

[20] Zipf, H.P. 1949. *Human behavior and the principleof least effort.* Cambridge, Massachusetts: Addison-Wesley.

[21] Luhn, H.P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* **2**(2): 159-165.

[22] Ahmad,F. 1995. A Malay Language Document Retrieval System: An Experimental Approach And Analysis. PhD thesis. University Kebangsaan Malaysia.

[23] Salton, G. and McGill, M.J. 1983. *Introduction to Modern Information Retrieval.* New York : McGraw-Hill.

[24] Sembok, T.M.T., Palasundram, K., Ali, N.M., Aidanismah, Y., Wook, T.M.T. 2003. Istilah Sains: A Malay-English Terminology Retrieval System Experiment Using Stemming and N-grms Approach on Malay Words, Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access, 6th International Conference on Asian Digital Libraries, ICADL 2003, Kuala Lumpur. Berlin: Springer.

[25] Hamzah, M.P., Sembok, T.M.T. 2004. Evaluating the Effectiveness of Various Similarity Measures on Malay Textual Documents. Proceedings of WSEAS'04, Canary Islands, Spain, December 17-19, 2004.

[26] M. Stanojević, S. Vraneš: Semantic Approach to Knowledge Processing, WSEAS Transactions on Information Science and Applications, Vol. 5, Issue 6, 2008.

[27] S. Pohorec, S., Verlič, M., and Zorman, M. 2009. Domain specific information retrieval system, Proceedings of the 13th WSEAS international conference on computers (part of the 13th WSEAS CSCC multiconference), July 2009, pp. 502-508.

[28] Kadir, R.A., Sembok, T.M.T, and Halimah, Z. 2006. Binding Skolem Clauses in Theorem Prover Resolution for Automated Hypothetical Question Answering. Proceedings of the 5th WSEAS International Conference on Computational Intelligence Man-Machine Systems and Cybernetics (CIMMACS`06).

APPENDIX A: SET OF MALAY QUERIES

1. *Kelahiran Nabi Isa.*
2. *Raja-raja yang beriman.*
3. *Tentang mata.*
4. *Apakah tanda-tanda hari Kiamat yang dinyatakan di dalam Al-Quran?*
5. *Apakah hadis atau ayat Al-Quran yang membuktikan kewujudan alam ini?*
6. *Sebutkan ayat Al-Quran yang menekankan kepada wajibnya seorang wanita itu menutup aurat.*
7. *Dari ayat/surah manakah boleh didapati berkenaan kepentingan ilmu?*
8. *Dari ayat/surah manakah boleh didapati tanggungjawab seorang anak kepada ibu bapanya?*
9. *Dari ayat/surah manakah boleh didapati tanggungjawab seorang abang/kakak terhadap adik-adiknya?*
10. *Saya ingin tahu ciri-ciri berpakaian yang dikehendaki di dalam Islam yang patut dipatuhi oleh umatnya.*
11. *Nyatakan ayat-ayat mana dalam Al-Quran yang menyatakan tentang kelebihan berpuasa/sembahyang dari segi kesihatan dan kewajipan.*
12. *Dalam surah apakah yang berkaitan dengan cerita nabi-nabi?*
13. *Dalam surah manakah yang banyak membicarakan tentang hudud?*
14. *Ingin dapatkan ayat yang mewajibkan sembahyang Jumaat.*
15. *Adakah terdapat pernyataan tentang kewujudan makhluk lain di tempat lain/planet lain selain dari di bumi dalam Al-Quran?*
16. *Surah-surah dan pada ayat ke berapa terdapat maklumat berkenaan sesuatu hukum contohnya berkenaan dengan hukum mencuri?*
17. *Saya adalah seorang lelaki. Saya ingin dalil Al-Quran yang menyatakan tentang perempuan-perempuan yang haram saya kahwini.*
18. *Saya inginkan pilihan ayat-ayat Al-Quran yang menyatakan kebesaran Allah melalui kejadian gunung-ganang atau laut untuk menyelesaikan album 'kebesaran Allah'.*
19. *Perkara-perkara serta hujah yang menunjukkan Al-Quran itu adalah satu mukjizat.*
20. *Capaian maklumat berkenaan rukun Islam iaitu kalimah syahadah.*
21. *Capaian maklumat berkenaan rukun Islam iaitu sembahyang.*
22. *Capaian maklumat berkenaan rukun Islam iaitu puasa.*
23. *Capaian maklumat berkenaan rukun Islam iaitu zakat.*
24. *Capaian maklumat berkenaan rukun Islam iaitu haji.*
25. *Sembahyang-sembahyang sunat.*
26. *Apakah perbezaan antara Islam dan Kristian dan Yahudi yang disebut dalam Al-Quran?*
27. *Adakah dinyatakan secara jelas perkaitan antara para Nabi dari segi persamaan, keturunan atau ciri-ciri seorang nabi atau rasul?*
28. *Maklumat-maklumat/ayat mengenai hari kiamat.*
29. *Ayat-ayat Al-Quran yang menceritakan tentang sejarah peperangan yang telah berlaku.*
30. *Ayat-ayat Al-Quran yang menerangkan bab perkahwinan.*
31. *Dalam surah manakah menceritakan kisah beberapa orang lelaki tertidur di dalam gua selama beratus-ratus tahun?*
32. *Senaraikan nama surah-surah yang menceritakan atau menggambarkan keadaan jannah dan neraka.*
33. *Maklumat berkenaan dengan menyucikan diri (bersuci).*
34. *Maklumat tentang tuntutan berperang pada jalan Allah.*
35. *Maklumat tentang tuntutan berdakwah.*
36. *Apa jua kejadian malapetaka di bumi Allah ini merupakan satu petunjuk daripada Allah. Bagaimana Al-Quran dapat membuktikannya (melalui isi kandungannya)?*

(Note : For the relevant judgement please refer to Ahmad [22])

APPENDIX B

Table-6. Best Average Recall-Precision Values of Various Experiments (refer to Keys for the meaning of abbreviations)

		Recall Values										
Exp	**Thr**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**Average**
bsksd	1.0	0.391484	0.328080	0.263509	0.231146	0.156648	0.140584	0.130282	0.072540	0.066624	0.052831	0.183373
tsksd	1.0	0.389386	0.320780	0.259413	0.218424	0.146872	0.135887	0.126141	0.068861	0.063680	0.052829	0.178227
brsksd	0.8	0.378653	0.309316	0.250050	0.209260	0.146191	0.134998	0.125308	0.068771	0.063618	0.052824	0.173899
tnknd	0.6	0.360302	0.270907	0.230304	0.206354	0.137955	0.095774	0.076317	0.039792	0.031956	0.028474	0.147814
brnknd	0.6	0.352555	0.249903	0.209216	0.189971	0.123364	0.112822	0.104299	0.057402	0.048723	0.040193	0.148845
tsknd	0.6	0.348821	0.247188	0.221274	0.199600	0.130944	0.096458	0.076618	0.039492	0.028743	0.025527	0.141466
bnknd	1.0	0.343802	0.277457	0.243808	0.208001	0.128378	0.091160	0.049507	0.034135	0.028543	0.025479	0.143027
bsknd	0.7	0.340093	0.244674	0.220026	0.192138	0.128481	0.093494	0.072450	0.037318	0.028448	0.025363	0.138249
NC		0.337437	0.275656	0.238384	0.203974	0.126217	0.088235	0.046583	0.035510	0.029375	0.026312	0.140768
brsknd	0.7	0.326174	0.238434	0.218717	0.193873	0.128521	0.094869	0.074654	0.037600	0.028635	0.025563	0.136704

Keys

Exp	**Exp**eriment/Method
b	**b**igram
t	**t**rigram
s	**s**temmed
n	**n**onstemmed
d	**d**ocument
Thrs	**Thr**eshold Values
k	**k**eywords
NC	**N**on-**C**onflation