Evaluating The Effectiveness Of Various Similarity Measures On Malay Textual Documents

MOHD POUZI HAMZAH[1] and TENGKU MOHD TENGKU SEMBOK[2]
[1]Department Of Computer Science, Kolej Universiti Sains & Teknologi Malaysia, 21030
MALAYSIA.

[2]Department Of Information Sciences, Universiti Kebangsaan Malaysia, 43600
MALAYSIA

Abstract. This paper analyzes the effect of various similarity measures namely inner product for un-weighted query terms, inner product for weighted query terms, cosine of the angle between query and document vectors and Euclidean distance. The study was motivated by the fact that many researchers especially in Malay document retrievals tend to use simple method of calculating similarity measure that is based on inner product for un-weighted query terms. This paper shows that Euclidean distance outperforms other similarity measures significantly. The results suggest that Euclidean distance should be used to improve performance of document retrieval systems.

Keywords: Information retrieval; Malay language; Vector space; Similarity measure; Retrieval effectiveness

1 Introduction

There are many document retrieval systems and there have been many advances in areas such as keyword retrieval, similar file retrieval, automatic document classification and document summarization [1,3,5,6,12,13]. Methods to find similarity between query and documents have long been recognized as having a significant effect on the performance of text based document retrieval systems.

In this study we investigate various methods of similarity measures in order to find the best method that can be used for Malay document retrieval systems.

Much recent retrieval system is based mainly on user criteria [3] such as type of output presentation, amount of user effort needed for search and level of coverage of the target collection. The most important measures are:

(a) ability of the system to retrieve wanted information
(b) ability of the system to reject unwanted information.

Several evaluation studies use test methodology based mainly on Recall value and Precision value that apply to a set of test similarities [7,8]. To generate Recall and Precision requires:

(a) differentiation between similar retrieved documents and similar documents that are not retrieved
(b) differentiation between similar documents that are relevant to input and similar documents that are not relevant to input.

The classification technique of documents are based on Vector-Space models [9,10] and Probabilistic models. These methods make it possible to retrieve and classify texts according to arbitrary databases without referring to systematic classified information.

2 Experimental Detail

In any experimental document retrieval system, test collection consists of document database, set of queries for the database and relevance judgments that are formulated based on the queries[2,8,11].

2.1 Test Collection

To date there is only one Malay test collection. This collection is compiled by Fatimah [4]. Fatimah has separated the Quranic document into 6236 documents according to verses in Quran. According to Salton [7,8] and Van Rijsbergen [11] documents in a collection must be independent from each other. However in current collection the documents are not independent. There are some verses (documents) that cannot exist without their prior verses.

2.1.1 New Collection

We make a new compilation of documents by separating the documents into independent sections. Instead of 6236 documents in Fatimah's collection, a new collection consists of 811 documents.

2.1.2 Query Statements

In this experiment, the query statements are taken from Fatimah's collection [4]. There are 36 natural language query statements in the collection.

2.1.3 Relevance Judgment

We make a new compilation of user-defined relevance judgment based on a new document collection.

2.2 Similarity Measure

Similarity between query and document is based on query terms vector and document terms vector. The weight of each term is given by the following equations:

$$W_{ij} = tf_{ij} \times idf_i \qquad (1)$$

where,

$$tf_{ij} = \frac{frek_{ij}}{frek_m} \qquad (2)$$

$frek_{ij}$ – no. of terms i in document j

$frek_m$ – total terms in document j

and,

$$idf_i = \ln\left(\frac{N}{n_i}\right) \qquad (3)$$

n_i – no. of documents where term i exist.

N – no of documents in a collection.

A weight for each query term is as follows [2]:

$$W_{iq} = (0.5 + 0.5 * tf_{iq}) \times idf_i \qquad (4)$$

Our study focuses on four different methods of finding similarity between query and document. The mathematical equation for each method is as follows:

(a) Inner product for un-weighted query terms

$$iprod = \sum\left(W_{ij} * W_{iq}\right) \qquad (5)$$

where W_{iq} = 1 if term i exists in a query or 0 otherwise.

(b) Inner product for weighted query terms

$$iprod = \sum\left(W_{ij} * W_{iq}\right) \qquad (6)$$

where W_{iq} is calculated using equation (4) above.

(c) Cosine of the angle

The Cosine of the angle between query vector and document vector is given by the following equation:

$$\cos\theta = \frac{\sum\left(W_{ij} * W_{iq}\right)}{\sqrt{\sum W_{ij}^{2}} * \sqrt{\sum W_{iq}^{2}}} \quad (7)$$

(d) Euclidean distance

The distance between two points in vector space is given by the following equation:

$$\text{distance} = \sqrt{\sum\left(W_{ij} - W_{iq}\right)^{2}} \quad (8)$$

2.3 Evaluation

There are many ways to evaluate document retrieval systems [11]. In our experiment we use precision at standard recall points and R-precision to compare effectiveness of the systems.

$$\text{Recall(R)}=\frac{\text{number of documents retrieved and relevant}}{\text{total relevant documents from collection}}$$

$$\text{Precision(P)}=\frac{\text{number of documents retrieved and relevant}}{\text{total documents retrieved from collection}}$$

To further compare effectiveness of the systems, we use R-precision that is the precision at the R-th position in the ranking of results for a query that has R relevant documents [2].

3 Result

The experimental results in table 1 to table 4 show standard recall (R), precision (P) for retrievals using inner product for un-weighted query terms, inner product for weighted query terms, cosine of the angle between query and document vectors and Euclidean distance.

Method 1 – Inner Product for un-weighted query terms
Method 2 – Inner Product for weighted query terms
Method 3 – Cosine of the angle
Method 4 – Euclidean Distance

Table 1
Precision at standard recall for method 1

Recall	Precision
0.1	0.31290
0.2	0.22981
0.3	0.16964
0.4	0.14928
0.5	0.14880
0.6	0.13052
0.7	0.11580
0.8	0.07756
0.9	0.05789
1.0	0.03379
Average	0.14260

Table 2
Precision at standard recall for method 2

Recall	Precision
0.1	0.37202
0.2	0.27268
0.3	0.24244
0.4	0.21182
0.5	0.19251
0.6	0.16339
0.7	0.14194
0.8	0.10354
0.9	0.07768
1.0	0.05356
Average	0.18316

Table 3
Precision at standard recall for method 3

Recall	Precision
0.1	0.37490
0.2	0.30985
0.3	0.27753
0.4	0.24182
0.5	0.23175
0.6	0.20123
0.7	0.18187
0.8	0.13276
0.9	0.10328
1.0	0.07853
Average	0.21335

Table 4
Precision at standard recall for method 4

Recall	Precision
0.1	0.39583
0.2	0.30824
0.3	0.27180
0.4	0.23302
0.5	0.20925
0.6	0.18583
0.7	0.16165
0.8	0.13527
0.9	0.11303
1.0	0.10159
Average	0.21155

Fig. 1 shows method 1 yield lowest precision at every recall point, followed by method 2. Method 3 and 4 produce better result and they perform equally good. At some recall points method 3 is better than method 4.

Analysis from table 5 shows that method 3 and 4 share 47.2% the same value of R-precision. Out of 36 queries, method 4 has higher R-precision for 15 queries compared to only 4 queries for method 3.

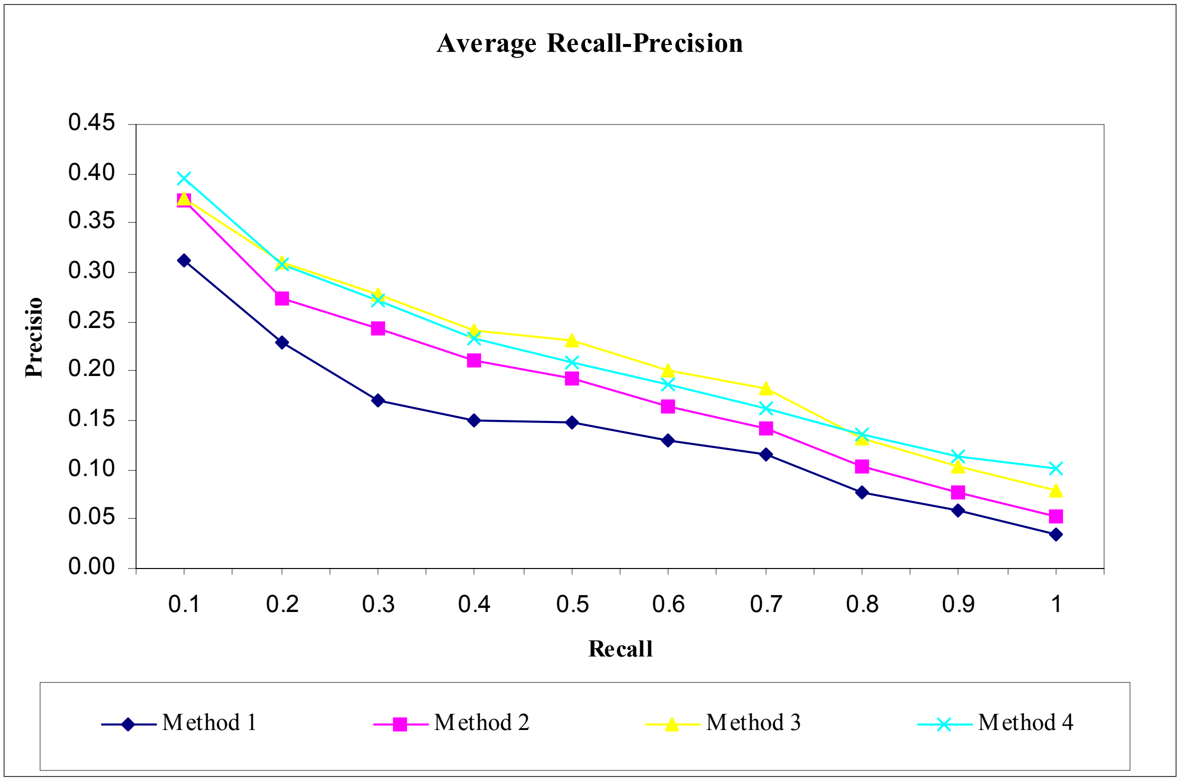

Fig.1. Average precision at standard recall for four methods of similarity measure

Table 5
R-precision for each query for four methods of similarity measure

Query	Method 1	Method 2	Method 3	Method 4
1	0.00000	0.14286	0.14286	0.14286
2	0.20000	0.20000	0.20000	0.20000
3	0.19231	0.19231	0.15385	0.23077
4	0.00000	0.00000	0.00000	0.00000
5	0.02778	0.02778	0.02778	0.11111
6	0.16667	0.33333	0.33333	0.33333
7	0.00000	0.00000	0.00000	0.00000
8	0.12500	0.12500	0.25000	0.25000
9	0.00000	0.00000	0.00000	0.00000
10	0.00000	0.00000	0.00000	0.00000
11	0.28125	0.31250	0.25000	0.25000
12	0.46364	0.45455	0.42727	0.26364
13	0.00000	0.00000	0.00000	0.00000
14	0.00000	1.00000	1.00000	1.00000
15	0.08511	0.08511	0.08511	0.17021
16	0.00000	0.00000	0.00000	0.33333
17	0.25000	0.25000	0.00000	0.25000
18	0.06250	0.06250	0.06250	0.12500
19	0.22222	0.16667	0.16667	0.16667
20	0.09091	0.09091	0.09091	0.09091
21	0.43590	0.48718	0.52564	0.46154
22	0.12500	0.12500	0.12500	0.12500
23	0.32759	0.46552	0.51724	0.53448
24	0.27273	0.54545	0.54545	0.63636
25	0.00000	0.00000	0.00000	0.00000
26	0.42012	0.42604	0.44379	0.49704
27	0.24242	0.25758	0.31818	0.37879
28	0.46875	0.43125	0.44375	0.45625
29	0.00000	0.00000	0.00000	0.06667
30	0.02703	0.08108	0.02703	0.08108
31	0.00000	0.00000	1.00000	0.00000
32	0.25000	0.19792	0.27083	0.33333
33	0.12500	0.12500	0.12500	0.12500
34	0.35417	0.45833	0.47917	0.43750
35	0.00000	0.00000	0.00000	0.19231
36	0.00000	0.00000	0.00000	0.00000

4 Conclusion

Many document retrieval systems use simple method to calculate similarity between query and document. This paper shows inner product for un-weighted query terms and inner product for weighted query terms produce poor result. Euclidean distance outperforms other three measures.

References:

[1] Atlam, E.S., Fuketa, M., Morita, K., & Aoe, J., Documents Similarity Measurement Using Field Association Terms, *Information Processing and Management Journal*, 39, 2003, pp. 809-824.

[2] Baeza-Yates, R & Ribeiro-Neto, B., *Modern Information Retrieval,* Addison-Wesley, New York, 1999.

[3] Croft, W. B., User-specified Domain Knowledge for Document Retrieval, *Proceedings Of The ACM Conference On Research And Development In Information Retrieval*, 1986, pp. 201-206.

[4] Fatimah A., *A Malay Language Document Retrieval System: An Experimental Approach And Analysis*, Ph.D Thesis, Universiti Kebangsaan Malaysia, 1995

[5] Fagan, J. L, *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods*, Ph.D. Thesis, Department of Computing Science, Cornell University, Ithica, New York, 1987

[6] Sanderson, M. ,Word Sense Disambiguation and Information Retrieval, *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* 1994, pp. 142-151, Springer-Verlag.

[7] Salton, G., A Blueprint For Automatic Indexing, *ACM SIGIR Forum 16*, 2 (Fall 1981), 1981, pp. 22-38.

[8] Salton, C.. and Lesk., M.E. Computer Evaluation Of Indexing And Text Processing, *Communication of the ACM*, Vol 15 No. 1 , 1986, pp. 6-36..

[9] Salton, G., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[10] Salton, G., Another Look At Automatic Text Retrieval Systems, *Communications of the ACM*, Vol 29 No. 7, 1986, pp. 648-656

[11] Van Rijsbergen, C.J. *Information Retrieval*, 2nd edition, Butterworth.,1979

[12] Zainab Abu Bakar, *Evaluation Of Retrieval Effectiveness Of Conflation Methods On Malay Documents*, Ph.D Thesis, Universiti Kebangsaan Malaysia, 1999.

[13] Zainab Abu Bakar & Nurazzah Abdul Rahman, Evaluating The Effectiveness Of Thesaurus And Stemming Methods In Retrieving Malay Translated Al-Quran Documents, *Proceeding Of 6th International Conference On Asian Digital Libraries*, 2003, pp. 653-662. Springer-verlag.