Improving the Precision of Al-Quran Retrieval Using Latent Semantic Indexing with Background Knowledge

Susanti
Department of Information Technology, STMIK-AMIK Riau, Pekanbaru, Riau
susanti07@gmail.com

Abstract

The aim of this study is to test the effectiveness of Al-Quran precision retrieval using LSI with background knowledge. The primary data used during the testing is the English translation of the Al-Quran where as the English translated hadith is used as the secondary data which acts as the background knowledge. SVD that is an LSI algorithm, indexes training data to be accessed by query. Experiments conducted are encircled around the two models i.e. LSI with the background knowledge and LSI without the background knowledge. The retrieval effectiveness is measured using the standard precision and recall measures.

Kata Kunci : LSI, Background Knowledge.

1. Introduction

Some studies of the expansion of information retrieval model were done to improve the ability of retrieval system in providing user information need. The model of information retrieval plays an important rule in giving the output wanted by user. Output as a document is relevant based on the system, but sometime is not relevant according to user, because the relevancy is decided by user.

The expansion of information retrival model is aimed to solve the problem in information retrival, such as synonym and polysemy. Polysemy is a factor that make the low retrieval precision while synonym tends to reduce the recall for retrival system [Deerwester et al, 1990]. The precision and recall method can be used to measured the retrival effectiveness.

One of the document that has problem in information retrival is Al Quran document. In this document, most of the studies used vector space model. LSI is a variant of the VSM which maps a high dimensional space into a low dimensional space, (Kumar and Srinivas, 2006). The LSI model is an expansion of VSM that may solve the synonym and polysemy problem in retrieval. The application of LSI model in Al Quran document is very limited. Some studies make improvisation of LSI model, by adding background knowledge. Therefore, this study was conducted to test the effectiveness of LSI with background knowledge of Al Quran document.

In this paper we demonstrate how LSI with background knowledge may improve the the retrieval precision of Al Quran document. Some previous studies had been reviewed to support this experiments.

2. Background and Related Research

2.1 LSI

LSI [Deerwester et al, 1990], is one of information retrieval model that tries to remove the problems lexical matching by using statistically derived conceptual indices instead of individual words for retrieval. LSI assumes that there is some underlying or latent structure in word usage that is partially obscured by variability in word choice [Barbara Rosario, 2000; Kontostathis and Pottenger, 2004].

According to Price and Jucas (2005), LSI is a robust dimensionality-reduction technique for the processing of textual data. LSI has SVD algorithm that based on math algebra. SVD algorithm is a method to decomposition matrix term document into three matrix, which their value can describe the relationship between the words in one dimension vector (Dumais, 1991; Gao and Jhang, 2004). The term document matrix detect the frequency of term in each document. For instant, the frequency of word i in document j in matrix term-document is A =[aij]. According to Gao and Jhang (2004),more than 99 %,

the value of aij is zero, because not all the words can be found in all documents. The formula of SVD is as followed :

$A = U \Sigma V^T$

A = Matrik term-document *(t,d)* , each of word frequency *(t)* is located in rows of each document coloumn *(d).*

U = Orthogonal matrix multiplies word by n=min(t.d) of vector coloumn, namely left singular vector of matrix A.

V = Orthogonal matrix multiplies document (d) by n = min (t,d), namely right singular vector of matrix A and n is the rank of A, (Dumais, 1994: Rosario, 2000)

Σ = As *singular value*, $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots\ldots \sigma_r \geq 0$.

V^T= Tranpose Matrix of V

SVD may truncate the data, due to its rank-k method that take SVD matrix with certain coloumn, based on the *k* value, in providing the relevant document. The query matching in LSI is not directly connected to term-document matrix. This method is different with vector space model, where the query is connected to the term-document matrix. LSI used the rank-k matrix, connected to query. The reduction of rank-k with the lowest value in dimension, according to Berry et al. (1995), may gives document that can be presented and remove noisy data from term-document matriz and *k* value is number of coloumn saved alter SVD (David and Ophir, 2004).

2.2 Background Knowledge

The meaning of background knowledge method is explained as mentioned bellow :

a. relevant textual background data allows for richer co-occurrences to be modeled properly. The background knowledge must close to the training data. If the background added have a connection with training data, so that the result will have precision retrival improvement. To be more useful in SVD process, a background set should contain similar data to most of the training set. (Zelikovitz dan Hirsh, 2001 : Zelikovit dan Marquez , 2005).

b. The background knowledge is not included in the list of relevant document, because it is an supporting data for training data. (Zelikovitz dan Hirsh, 2001).

2.3 Related Research

The similar study had been conducted by Zelikovittz and Hirsh (2001). They used a LSI method by giving background knowledge for text classification, with the conclusion that background knowledge may improve the document retrival precision of LSI model. In 2005, Zelikovitz and Marquez studied an evaluation of background knowledge for LSI classification. They may concluded text classification experiment that to be most useful in the singular value decomposition process, a background set should contain data that is similar to much of the training set.

3. Empirical Result

3.1 Data Set

Al Quran document in English version, use hadith as a background knowledge. Index Al Quran book is aguide to conduct this research, because in that index have contained surah with a relevant topic. This topic can be treated as query in retrieval system. A total of 10 queries were run on both experiments. The training document in this study, was based on the given query. Surah that contain relevant word with query ia as a training document, while hadith related to that surah is as a background knowledge.

3.2 Result

The evaluation of the IR systems is usually done with the standard measures of precision (P) and recall (R), where:

$$P = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

$$R = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$

The result of this study is presented in table 1 and figure 1.

Table 1. The precision average of 2 experiments

RECALL (%)	PRESICION (%)	
	LSI	LSI BK
10	48.69	55.25
20	44.44	51.39
30	45.68	49.59
40	42.63	46.07
50	39.63	37.77
60	31.43	36.61
70	27.41	32.43
80	26.52	29.29
90	26.62	27.62
100	25.42	28.31
Average	35.85	39.43
Percentage of precision improvement (%)		**3.58**

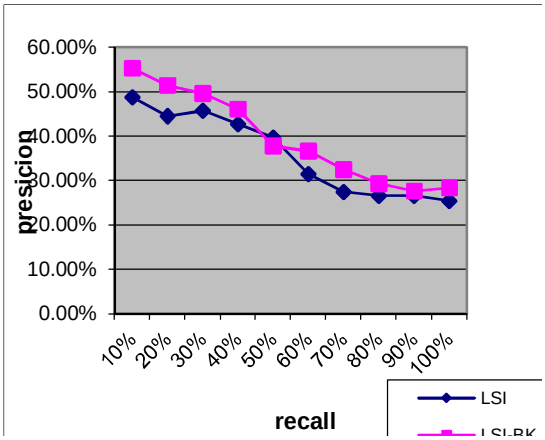

Figure 1 : LSI and LSI-Bk for 10 queries

As seen in table 1, there is an the improvement of LSI precision background knowledge, that elarly found in recall percentage 10 an 20 %, with improvement 6.55% and 6.95 % respectively. But in recall 50 %, the precision decreased to 1.86 %. According to Latmas et al. (2002): Sembok (2007) , 85 % user just checked the document in the first page, that means the precision measurement is more important than recall measurement.

Based on the result shown in Table 1 and figure 1 above, it's clearly seen that LSI with background knowledge give the higher precision improvement (3.58%).

4. Conclusion

The experiment LSI with and without background knowledge have been done into document's al-Quran. The comparison result of these two experiments proved that there is an improvement of retrieval precision by LSI model with background knowledge in Al-Quran retrieval.

References

[1] David, A. G. ,& Ophir,F. 2004. Information Retrieval algorithms and heuristic. second edition. Netherlands. Springer**.**

[2] Deerwester, S, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391-407.

[3] Gao,Jing & Zhang,Jun. 2004. Clustered SVD strategies in latent semantic indexing. Journal of Information Processing and Management 41 (2005) 1051–1063.

[4] Kumar, C.A & Srinivas.S. 2006. latent semantic indexing using eigenvalue analysis for efficient information retrieval. Journal appl. math. comput. sci., vol. 16, no. 4, 551–558.

[5] Kontostathis,April and Pottenger,W.M. Preprint submitted to Elsevier Science 30 June 2004, A Framework for Understanding Latent Semantic Indexing (LSI) Performance.

[6] Price, R.J and Zukas, A,E. 2005. Application of Latent Semantic Indexing to Processing of Noisy Text. Eds.: ISI 2005, LNCS 3495, pp. 602 – 603.Springer-Verlag. Berlin Heidelberg

[7] Sembok, T.M.T. 2007. Bahasa, Kecerdasan dan makna sekitar capaian maklumat, Syarahan perdana, Universiti Kebangsaan Malaysia.

[8] Zelikovitz, Sarah and Marquez, Finella.2005. Evaluation of Background Knowledge for Latent Semantic Indexing Classification. Journal of American Association for Artificial Intelligence.