Pertanika J. Sci. & Technol. 25 (S): 99 - 108 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

The Implementation of Latent Semantic Indexing on Knowledge Retrieval Process in Knowledge Management System Development

Novi Sofia Fitriasari*, Rani Megasari and Arum Yuniarsih

Departemen Pendidikan Ilmu Komputer, Universitas Pendidikan Indonesia, Jl. Dr.Setiabudhi Nomor 229 Bandung 40154, Indonesia

ABSTRACT

This study examines Latent Semantic indexing (LSI) using Singular Value Decomposition (SVD) in the knowledge retrieval process, namely indexing Indonesian text. There are three stages in this process: (1) text processing, which consists of tokenisation, filtering, and stemming process, (2) developing LSI using SVD and (3) evaluating and measuring performance. The result showed Mean Average Precision around 77.90% on scenario matrix dimension 120 and average precision for first retrieval around 83.33% on scenario matrix dimension 90.

Keywords: Knowledge Retrieval, LSI, SVD

INTRODUCTION

Knowledge management is very important for the success of any organisation (Fitriasari, Sofia, Suputra, Wirya, & Dini, 2012). In an organization, knowledge resides in various forms, such as documents, electronic databases, codified human knowledge stored in expert system, documented organisational procedures and processes and tacit knowledge acquired by individuals and networks of individuals (Tan, Teo, Tan, & Wei, 1998)

Knowledge management has four "knowledge processes", namely: (1) creation, (2) storage and retrieval, (3) transfer, and (4) application. Each process is not monolithic but are interlinked and affect one another (Alavi & Leidner, 2001).

Storage, organisation and retrieval of memory or knowledge of an organisation are important aspects for a knowledge management system that must be applied effectively (Babu, Vardhan, & Kuar, 2012).

ARTICLE INFO

Article history:
Received: 12 January 2017
Accepted: 02 October 2017

E-mail addresses:
novisofia@upi.edu (Novi Sofia Fitriasari),
megasari@upi.edu (Rani Megasari),
arum.yuniarsih@student.upi.edu (Arum Yuniarsih)
*Corresponding Author

ISSN: 0128-7680

Thus, knowledge retrieval mechanism that is easy to remember and to use is essential in the strategy of Knowledge Management in an organisation (Tan et al., 1998). In the absence of an effective retrieval process, the knowledge that has been created and stored as an organisational knowledge cannot be transferred to the user actor and the knowledge application process would not be efficient.

Referring to this framework, the retrieval process can be seen as the link between the process of creation and transfer of knowledge to application. The ISO/IEC 2382-1 defines information retrieval as acts, methods and procedures carried out to retrieve stored data in order to provide information on the needed subject. The main purpose of information retrieval mechanism is to meet the users' information needs by retrieving all documents that may be relevant and, at the same time, leaving irrelevant documents on the list of search result. Such system uses a heuristic function to obtain documents relevant to the user's query. Some of the methods used in information retrieval mechanisms are Boolean, Probabilistic, Vector Spaces, Fuzzy, P-Norm and Network Inference.

However, the process of retrieving organisational knowledge is not easy because the knowledge is stored in various forms and components and can be both structured and unstructured. Search and retrieval using a regular query cannot find knowledge that is relevant to the needs of users. Moreover, technically, if the process of knowledge retrieval is performed using regular search that only relies on keywords or regex rules, in order to obtain relevant results, the number of rules that should be used will be very large and difficult to monitor (non-scalable) along with increasing scope of knowledge stored.

In the search for the most effective way to retrieve information, it is important to consider existing models since using available models and methods is very helpful in the decision-making process and has an important role in knowledge management (Babu, Vardhan, & Kuar, 2012). Latent Semantic Indexing, as proposed by Phadnis et al., (2014), has its advantages as a means of retrieving information, in that, it retrieves documents efficiently. It is a method of retrieval and indexing by means of a mathematical technique called Singular Value Decomposition (SVD) (Praks, Dvorsky, & Snase, 2003). Several research groups have used LSI for knowledge retrieval (see Table 1).

Table 1
Implementation of LSI based on references

Implementation	Reference
Query processing for medicine document	(Guo, Berry, Thompson, & Bailin, 2003)
The Arabic document retrieval	(Muqthader, 2007)
Nursing Knowledge Classification	(Chinchanikar, 2009)
Document retrieval for a very large data set	(Zaman, 2010)
Chinese Information Retrieval	(Luo, 2013)
Topic Modelling for Knowledge Discovery	(Xu, Li, & Craswell, 2013)
Idenditification of domain term in the source code	(Sharma, 2014)
Semantic search technology	(Li, Bhatia, & Cao, 2015)
Automating Traceability Link	(Mounika & Babu, 2016)

The SVD is usually used in a variety of applications, such as latent semantic indexing, collaborative filtering and general expression analysis (Rajamanickam, 2009). It is the most powerful decomposition algorithm in the LSI process in terms of the number of documents returned (Jaber Amira & Milligan, 2012).

The use of SVD for latent semantic indexing has been done by several research groups, as seen in Table 2.

Table 2
Function of SVD based on references

Function	References
Capturing the geometric structures of motion data	(Prabhakaran, Nadin, & Khan, 2006)
Band and Sparse matric	(Rajamanickam, 2009)
Looking up document	(Da, 2015)
Clusters inter document	(Zhang, Xiao, Li, & Zhang, 2016).

However, as seen in Table 1, LSI is the most popular knowledge retrieval mechanism in which the indexing system is predominantly based on English. There are limited sources of its implementation with its index based on a mix of English and non-English language. Therefore, this study discusses the implementation of LSI with SVD with concerns over indexing texts in Indonesian language.

METHODS

Text Processing and Index Development

The study examined selected theses written in Indonesian language submitted to Department of Computer Science Education of Indonesia University of Education (UPI). However, as the theses were related to computer science, English terms were widely used.

The document collection stage went through several phases before data could be used to develop the index. Those stages were tokenisation, filtering, and stemming. Tokenisation is the process of separating words contained in sentences or paragraphs which would later be processed in text analysis. A token is separated by spaces and punctuation. Each token then went through the stage of case folding in which a token was converted to lowercase to facilitate the next text processing. Tokens obtained from the tokenisation process were then filtered and tokens/terms listed in stop-word list went through the removal process

After being filtered, the tokens were stemmed. Stemming is a process of tracing the words in a document to their root by following certain rules. As an illustration, the terms "*bersama*" (together), "*kebersamaan*" (togetherness), "*menyamai*" (equal), are stemmed to their root, which is "*sama*" (same). There are several algorithms available for stemming Indonesian language, such as Nazief-Andriani's algorithm, Arifin-Setiono's algorithms, and Mustafa-Idris's algorithms. Stemming in this study employed Nazief-Andriani algorithms on program implementation.

From the results of the text processing, 2868 unique terms and term-document matrix sized 139x2868 were obtained. This matrix was further processed in matrix decomposition phase using SVD (Singular Value Decomposition) technique, which was followed by the development of latent semantic index. The matrix decomposition was reduced into a matrix of r-dimensional space. R value was determined by the researcher.

Evaluation *Scenario*

The system was evaluated using multiple evaluation scenarios. The case tested in this study was to determine the best r parameter in building sub-matrix.

In the early stage, the results of singular value decomposition of matrix A, from the collection of research documents, are expressed as follows (Zhang, Xiao, Li, & Zhang, 2016):

$$A = \quad U \quad \sigma \quad V^T$$

$$A = [u_1, u_2, u_3, \ldots, u_k] \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \ldots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_k \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_k^T \end{bmatrix}$$

with k as the number of singular values of A(σ1, σ 2,….,k). Then, from k, r as the largest singular value was selected, which is $\sigma 1 \geq \sigma 2 \geq \ldots . \geq \sigma r > 0$, with $r<k$. After the value of r was determined, the matrix derived from the decomposition was reduced into r-dimensional space.

The value of k (or the number of document collections) in this research is 139. Experiments were conducted on the value of r by applying systematic sampling technique. The members of R sample were selected by multiplication of 10, from which 13 samples were selected for the value of r, namely 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 130. Each r value was then evaluated to find the value of r that gave the maximum value of MAP (Mean Average Precision).

RESULTS AND DISCUSSIONS

The effectiveness and efficiency of information retrieval are affected by the quality of its index. The selection of appropriate values of r parameter will affect the index built from the collection of documents. Indexing distinguishes a document from others in the collection. A small size index can give poor results and could neglect relevant documents. Yet, an index with large size results in the discovery of documents that are not relevant and reduces the search speed.

Best r parameter is determined by selecting a scenario that provides the highest MAP. Not only will the highest MAP enable relevant documents to be found more quickly and precisely, it will also affect the ranking of documents. This is because changes in the size of matrix dimensions influence the calculation of cosine similarity between the query and the documents.

To get the value of MAP for each scenario r parameter, the calculation on precision value at each point of retrieval for each scenario was done. The results of the calculation precision in the range of scenario r with parameter of 10 to 60 are shown in Figure 1. The results of the

calculation on precision of scenarios of r = 70 to 130 are shown in Figure 2. Precision and MAP are expressed as follows (Manning, Raghavan, & Schutze, 2009):

$$\boldsymbol{Precision} = \frac{\#(\boldsymbol{relevant\ items\ retrieved})}{\#(\boldsymbol{retrieved\ items})}$$

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{JK})$$

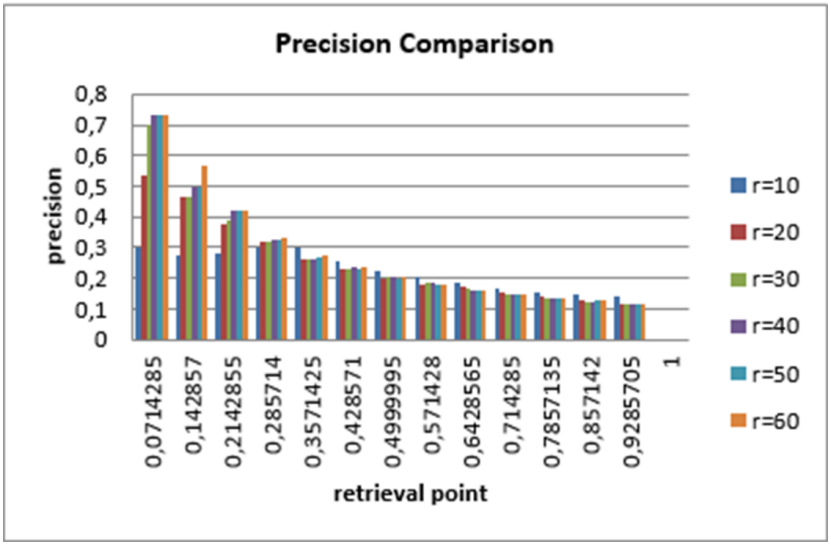

Figure 1. Comparison of Precision value for scenarios r = 10-60

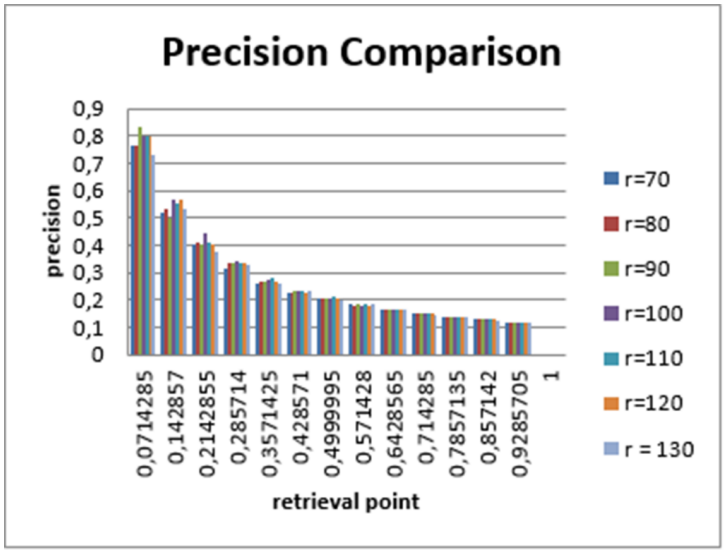

Figure 2. Comparison of Precision value for scenarios r = 70-130

The test results of all scenarios show r scenario with parameter value of 120 provides good precision value and the highest MAP is 77.90 %, as shown in Table 3

Table 3
Comparison of MAP value in each scenario the value of r

MAP	R Parameter Value
10	0.33503824
20	0.58148742
30	0.67072885
40	0.70011206
50	0.70313966
60	0.73573703
70	0.72484149
80	0.74624535
90	0.74422495
100	0.77728778
110	0.76744631
120	0.77905532
130	0.74130575

The results of the study show that latent semantic indexing is possible. It is proven by the discovery of almost all relevant documents at the first retrieval point. Some documents that do not contain the keyword in the title but contain related methods and is relevant to the query are also detected by the system. This is influenced by the application of LSI which can determine the semantic relationships between words in a document, in the term-document matrix used. Therefore, the quality of the built index will significantly affect the performance of the search. Good index will also lead to good search results.

However, there are some shortcomings, which are mainly related to the index used. The index quality is determined by the quality of term at the text processing stage.

Tokenisation uses a primitive separation technique which separates words with spaces and punctuation as references. Such a method has its disadvantages. As a result, there are irrelevant terms unfiltered and are listed in the list of terms.

For examples:

"*dianggapakuratdalampengenalanwajahadalahjaringan*"

"*Dibutuhkan*"

"*Desainperancangan*"

It occurs because as computer is not like human brain, the tokeniser is unable to recognise whether a group of characters that it separates is a meaningful or not. If an author miss-typed some words and they are not checked at the pre-data processing, the words will be taken into the list of terms and influence the size of the matrix and the process of index development.

In addition, for phrases in Indonesian, such as "*rumahsakit*" (hospital) and "*perguruantinggi*" (university), the tokeniser will separate them into distinct words, namely "*rumah*" (house), "*sakit*" (ill), "*perguruan*" (school) and "*tinggi*" (high). In Indonesian, such phrases form a single phrase. Separating "*rumahsakit*" into "*rumah*" and "*sakit*" changes the meaning of the phrase.

Moreover, as a result of the conversion of the letters into lowercase, some terms such as "*Budi*" (name of a person) might lose its meaning as it would be referred to as the same like in the phrase "*budipekerti*" (good personal characters).

In terms of filtering, stop words list that is used is that of standard Indonesian and have not been adjusted to the existing collection of documents. Yet, a word that appears too often is considered as unimportant and included in the list of stop-words.

In terms of stemming, the algorithm used is a stemming algorithm for Indonesian. The drawbacks of stemming used in this system are as follows:

a. In the document collection used in the study, many English terms are not affected by the stemming algorithm. English terms remained because many important terms in English are used in the field of computer science. Hence, there are some English terms left unfiltered and included in the term-document matrix processing.

b. The stemming process allows two phrases whose meanings are similar to be much different. For example, the words "*depolitisasi*" (depoliticisation) and "*politisasi*" (politicisation), which have similar meaning, will be stemmed into token "*depolitis*" and "*politis*", which have different meanings.

c. As the stemmer applies the rules on each term/token passing through text processing, many terms may lose their intended meaning. Terms such as "*wartawan*" (journalist), "*karyawan*" (employee), "*peragawan*" (male model) will be converted via a stemming process into "*warta*" (news), "*karya*" (works) and "*peraga*" (model). Phrases that have name meaning such as "*gunawan*" and "*setiawan*" would also change to "*guna*" (use) and "*setia*" (loyal), which deviate from their original meaning.

Therefore, it can be concluded that text processing performed in this study can process terms contained in the collection of documents However, for collection of documents with larger size and more diverse types of documents, further related research needs to be conducted.

CONCLUSION

The LSI module can be used on a collection of documents to index and to rank relevant documents. R values that is ideal in the implementation of LSI on the collection of documents in this study is 120, since the value of r = 120 gives MAP of 77.90% and with an average precision of 80

Based on the evaluation and performance measurement, latent semantic indexing method gives a fairly good performance, with a mean average precision of 77.90% in value scenarios r = 120 and average precision for first retrieval of 83.33% in the first scenario r = 90,

In addition, based on the identified shortcomings, the importance of an effective means of indexing texts, particularly those from non-English language, is inevitable. Hence, further research on the development of indexing process are essential.

REFERENCES

Alavi, M. & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Babu, K. V. S. N. J., Harshavardhan, T., & Kumar, A. J. S. (2012). The role of information retrieval in knowledge management. *International Journal of Social Science and Interdisciplinary Research, 1*(10), 212-226.

Chinchanikar, S. V. (2009). *Automated nursing knowledge classification using indexing*. M. S. (Unpublished doctoral thesis), Florida Atlantic University, USA.

Da, N. T. (2016). An approach to look up documents in a library using singular value decomposition. *International Journal of Computer Science and Network Security (IJCSNS), 16*(5), 123.

Fitriasari, Sofia, N., Suputra, Wirya, I. M. & Dini, H. (2012). Sistem manajemen pengetahuan PT pos Indonesia. *Prooceeding Konferensi National Sistem Informasi.* Bali, Indoesia: STIMIK STIKOM Bali.

Guo, D., Berry, M. W., Thompson, B. B. & Bailin, S. (2003). Knowledge-enhanced latent semantic indexing. *Information Retrieval, 6*(2), 225-250.

Jaber, T., Amira, A. & Milligan, P. (2012). Enhanced approach for latent semantic indexing using wavelet transform. *IET Image Processing, 6*(9), 1236-1245.

Prabhakaran, B., Nadin, M., & Khan, L. (2006). Efficient 3D Motion Pattern Retrieval in Large Motion Capture Databases.

Li, W., Bhatia, V. & Cao, K. (2015). Intelligent polar cyberinfrastructure: Enabling semantic search in geospatial metadata catalogue to support polar data discovery. *Earth Science Informatics, 8*(1), 111-123.

Luo, Y. (2013). An improved semidiscrete matrix decomposition and its application in Chinese information retrieval. *Applied Mechanics and Materials, 241*, 3121-3124.

Manning, C. D., Raghavan, P. & Schutze, H. (2009). *An introduction to information retrieval.* Cambridge: Cambridge University Press.

Mounika, D., & Babu, K. D. (2016). Automating traceability link recovery using information retrieval. *i-Manager's Journal on Software Engineering, 11*(1), 13.

Muqthadeer, A. M. (2007). *Arabic document retrieval using latent semantic indexing* (Published doctoral thesis), King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia.

Phadnis, N., & Gadge, J. (2014). Framework for document retrieval using latent semantic indexing. *International Journal of Computer Applications, 94*(14).

Praks, P., Dvorský, J. & Snášel, V. (2003, July). Latent semantic indexing for image retrieval systems. Proceeding of In *SIAM Linear Algebra. Philadelphia, USA: International Linear Algebra Society (ILAS)*.

Rajamanickam, S. (2009). *Efficient algorithms for sparse singular value decomposition*. University of Florida, Florida.

Sharma, K., & Brar, A. S. (2014). Visualizing the software system towards identifying the topic from source code using set mantic clustering. *International Journal of Advanced Computer Research, 4*(1), 350

Tan, S., Teo, H. H., Tan, B. & Wei, K. K. (1998). Developing a preliminary framework for knowledge management in organizations. *Proceeding of AMCIS 1998*. San Diego, USA: AMCIS.

Xu, J., Li, H. & Craswell, N. (2013). Microsoft Corporation; Patent issued for regularized latent semantic indexing for topic modeling. *Journal of Engineering*.

Zaman, A. N. K. (2010). *Study of document retrieval using latent semantic indexing (lsi) on a very large data set* (Published doctoral thesis), The University of Northern British Columbia, Canada.

Zhang, W., Xiao, F., Li, B. & Zhang, S. (2016). Using SVD on clusters to improve precision of interdocument similarity measure. *Computational Intelligence and Neuroscience*, 2016.