Information retrieval system in text-based skripsi document search file using vector space model method

Mochammad Abdul Azis¹, Abdul Hamid², Ahmad Fauzi³, Yudhistira⁴, Eko Yulianto⁵, Verry Riyanto⁶, Ridwansyah⁷, Sfenrianto⁸

¹,²,³,⁴,⁵ Information System, Bina Sarana Informatika University, Jakarta, Indonesia.
⁶,⁷ Departement of Informatic Technology, STMIK Nusa Mandiri, Jakarta, Indonesia.
⁸ Information Systems Management Department, BINUS Graduate Program–Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia.

E-mail: ¹mochammad.mmz@nusamandiri.ac.id, ²hamid.adh@bsi.ac.id,
³ahmad.aau@bsi.ac.id, ⁴yudhistira.yht@bsi.ac.id, ⁵eko.eui@bsi.ac.id,
⁶verry.vry@nusamandiri.ac.id, ⁷rdwansyah@gmail.co, ⁸sfenrianto@binus.edu

Abstract. Speed and simplicity in the process of finding documents and information have become a necessity in the campus library. The speed in finding a skripsi document that is currently still very long to find the data and more simple in displaying what you are looking for then we aim to search and find skripsi documents in a digital library at a private university in Jakarta more quickly and display what which is searched precisely. An approach is needed using the VSM method that encourages without complete keywords, and to see the effectiveness of research searches. The results of the VSM method can be applied for the development of prototype applications. The prototype used to facilitate the process of searching or finding the documents and information needed is fast and simple.

1. Introduction

In the digital search, the information era is used for many purposes in supporting information retrieval activities using information retrieval systems, the information sought from various content such as text, images, video, audio and so on for user needs.

This information is meaningless if the relevant information cannot be recovered for the needs of users, one of them is the library information application. Therefore, college libraries require information retrieval systems or digital library services that provide instant access to skripsi collections/documents through method keyword search, writing, and title.

The author tries to apply the retrieval information method to one of the private tertiary education libraries in Jakarta in order to provide faster and more accurate information in the skripsi document search process, so that students can do searches without needing to complete the keyword in detail, Hana students need to type the keyword in search of skripsi documents and will be displayed in full.

In this study, the author uses the vector space model method which is one of the retrieval information methods that aims to simplify the retrieval process in text-based documentation.

The implementation of the vector space model can be felt and enjoyed in various scientific fields, one of which is to develop document indexing schemes that increase the effectiveness of taking free text medical documents [1]. With documents and queries represented as vectors and angles between the two [2], vectors are calculated using the consinus function in common with the effectiveness of VSM which
largely depends on the weighting of the VSM [3] and allows the calculation results to be ranked according to the measure of similarity [4].

The purpose of this study is to implement a retrieval system model on library applications at private universities in Jakarta to trace documents in a repository based on document searches entered by students so that documents will be found similar to the keywords students have searched for, as well as ranking/order of documents which is found based on the similarity value.

2. Materials and methods

2.1 Information Retrieval System

Information retrieval system shows the process in the form of information retrieval needed [5]. Information retrieval (IR) is generally associated with knowledge-based information retrieval and retrieval [6] information retrieval system (IR).

Information Retrieval System finds information that is usually in the form of documents from unstructured data in the form of text to meet the information needs of a large set of data that is generally stored in a computer database [7].

The Information Retrieval System application has been used in many fields such as medicine, companies and so on. One application of the Information Retrieval System is a search engine that can be applied in various fields. On search engines with Information Retrieval Systems users can enter queries that are free in the sense of query words that correspond to human language and the system can find documents that match queries written by users.

The working principle of the Information Retrieval System is if there is a collection of documents and the user who formulates the question (request or request). The answer to the question is a series of relevant documents and discard irrelevant documents [8].

The Information Retrieval System will take one of these possibilities. Information Retrieval System is divided into two main components, namely, the indexing system to produce a system database and retrieval is a combination of user interface and look-up-table. The Information Retrieval System is designed to find documents or information needed by the user.

The Information Retrieval System aims to answer the information needs of users with information sources available in the following conditions [8]:
1. Presenting a set of ideas in a document using a set of concepts.
2. There are several users who need ideas, but cannot identify and find them well.
3. Information Retrieval System aims to bring together the ideas put forward by the author in documents with user information needs expressed in the form of keyword queries/search terms.

The main function of the Information Retrieval System [8]
1. Identify sources of information that are relevant to the interests of the targeted user community
2. Analyze the content of information sources (documents)
3. Representing the contents of the source of information in a certain way that allows it to be met with user questions
4. Represent user questions (queries) in a certain way that allows the source of information to be found in the database.
5. Meeting search statements with data stored in the database
6. Find back relevant information
7. Improve system performance based on feedback provided by the user.

2.2 Information Retrieval System Architecture

The Information Retrieval System process as in Figure 1 uses a simple architecture. Before doing the retrieval process, database definitions are needed. Next follows the stages of the process; The documents to be used, the operations that will be used in the search, and the text processing model [9].
2.3 Process Calculation of VSM

Vector Space Model (VSM) is a method to see the level of proximity or similarity (similarity) term by weighting the term. Documents are viewed as a vector that has a magnitude (distance) and direction (direction). In the Vector Space Model, a term is represented by a dimension of vector space. The relevance of a document to a query is based on the similarity between document vectors and query vectors [9].

VSM provides a partial matching framework is possible. This is achieved by setting non-binary weights for index terms in queries and documents. The term weight is finally used to calculate the level of similarity between each document stored in the system and user requests. Documents taken are sorted in a sequence that has similarities, the vector model takes into consideration the documents that are relevant to the user's request. The result is a collection of documents that are taken far more accurately (in the sense that they are in accordance with the information needed by the user). A \( dj \) document and a query \( q \) are represented as \( t \)-dimensional vectors as shown in figure 2.

![Figure 1. The Process of Retrieving Information [6]](image)

![Figure 2. The Cosines of is adopted as sim \( dj, q \) [6]](image)

![Figure 3. Matrix term-document [6]](image)
In VSM the document collection is represented as a matrix of term-document (or matrix term frequency). Each cell in the matrix corresponds to the weight given from a term in the specified document. A zero value means that the term is not in the document. Figure 3 shows the term-document matrix with \( n \) documents and \( t \) terms.

The process of calculating VSM is through the stages of calculating term frequency (Inf), term frequency Inverse Document Frequency (TF-IDF), query and document distances, measurement of query document similarity (inner product), and Cosine Similarity measurement (calculating cosine values the angle between two vectors).

Through VSM and TF-IDF weighting, you will get a numerical representation of documents so that you can calculate the closeness between documents. The closer the two vectors are in a VSM, the more like the two documents represented by the two vectors. The similarity between documents can be calculated using a similarity measure function. This size allows ranking documents according to their similarity or relevance to the query. Cosine Similarity or \( \text{Sim} (q, d_j) \) is used to evaluate the level of similarity or similarity of the document \( (d_j) \) relating to query \( (q) \) as a correlation between vectors \( d_j \) and \( q \). This correlation can be measured, with equation (1).

\[
\text{Sim}(q, d_j) = \frac{q \cdot d_j}{|q| \cdot |d_j|} = \frac{\sum_{i=1}^{n} W_{iq} \cdot W_{ij}}{\sqrt{\sum_{j=1}^{t} (W_{iq})^2} \cdot \sqrt{\sum_{i=1}^{n} (W_{ij})^2}} \quad \text{……………………… (1)}
\]

3. Results and Discussion

This designed system is used by private universities in Jakarta which already have student research skripsi data in digital form and stored in the repository.nusamandiri.ac.id. The data collected in the form of primary data with student skripsi data for 2015-2018 consisted of 800 attribute parameters with five input attributes used such as title, supervisor, student and year.

![Figure 4. The flow of the IRS System with VSM](image)

This system is designed by taking a repository database that has been available at repository.nusamandiri.ac.id at private universities in Jakarta.

Some things that must be considered within the boundaries of this study are as follows:

a) The system designed is an information retrieval system with documents in pdf format and
Indonesian language text.
b) The document is a document in the form of a text/paper.
c) The main data is a database that is registered at the repository.nusamandiri.ac.id

d) The system will read/retrieve document collections in the repository.
e) The search option that is made is based on the title, abstract and overall contents of the document.

The system description is designed using VSM and the process stages are seen in the system flow in Figure 4. The repository is a database whose documents will be retrieved and traced and the index file is a document store after the cluster is done according to the options. This application is designed with two users, the admin side and the user. Draft explanation as shown in Figures 5 and 6. The admin will do an indexing process wherein this process will be parsed and stemming the document, which after extracting it into format.txt is then grouped.

![Figure 5. Indexing Process (Admin)](image1)

![Figure 6. Indexing Process (User)](image2)
Next from the user side where the user only needs to enter the query to start the search for document searches. The document entered in the system is the document in the repository and the document is extracted into a .txt form as shown in Figure 7. The extraction process is done only by executing the extract file button in the application.

![Figure 7. Document Extraction](image)

For the next process that is still the admin right is to enter stopword data, and the interface for entering this stopword is shown in Figure 3. Entering the stopword can be done by uploading the file with the format. Txt and with a maximum size of 1024KB.

![Figure 8. The Stopword Upload Interface](image)

![Figure 9. Interface input query](image)
The first interface for the user is the column for entering the query as a keyword to start the search, which is shown in Figure 9 label a. The query entered is in the form of sentences and not files. The next step after entering the query, the user can choose the search option consisting of 4 (four) options as described in Figure 9 label b.

Figure 10 is the result that will be given to the user, while the documents have been ranked according to their similarity value to the query. The final result given to the user is the order of documents according to the similarity value to the user query, as described in Figure 16. The most similar document is P8 with a similarity value of 6.5%.

4. Conclusion
This system design can be used as a tool to get document references that are similar to the entered query so that the user can estimate the extent to which the research will be done. The prototype still has a system error when the user's choice option is content, this is possible because of limited hardware and time limit execution of the programming language used.

For the development of research, if you are going to use the Porter algorithm, you should add a complementary table to the Indonesian word so that all words can be recognized by the system. The collection of documents used should also be in a diverse format and will be maximized if using a wider database so that applications can be used together between institutions.

References
[1] Mao, W., Model ruang vektor berbasis frase untuk pengambilan otomatis medis teks bebas dokumen. Rekayasa Data & Pengetahuan, Volume 61 Edisi 1, 2007.
[2] Singh, V. K., Singh, V. K., Vishwavidyalaya, G., Programmer, S. A., & Vishwavidyalaya, G. G., Vector Space Model: An Information Retrieval, 141–143, 2015.
[3] Harcourt, P., & Japheth, R. B., Application of Vector Space Model to Query Ranking and Information Retrieval, 6(5), 42–47, 2016.
[4] Singh, J. N., Analysis of Vector Space Model in Information Retrieval, 14–18, 2012.
[5] Zhou, H., Liu, BW, & Liu, J., Penelitian tentang mekanisme pencarian informasi berbasis pada label ontologi. Procedia Engineering, 29, 4259-4266, 2012.
[6] Sharma, M., & Patel, R. Sebuah Survei Model Pengambilan Informasi, Teknik Dan Aplikasi. International Journal of Emerging Technology dan Advanced Engineering, 3 (11), 542–545, 2013.
[7] Manning, C., Raghavan, P., Pengantar Pengambilan Informasi, Stanford. AMERIKA SERIKAT, 2007.
[8] Salton, G., Pemrosesan Teks Otomatis, Transformasi, Analisis, dan Pengambilan informasi dengan komputer. Addison - Wesly Publishing Company, Inc. AS, 1989.
[9] A. Baeza , A. Pizza., Alberto Campo Baeza. Construction projects, 1989.