International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

http://ijpsat.ijsht-journals.org
Vol. 12 No. 2 January 2018, pp. 103-107

Vector Space Models on Data Search Applications Based on Information Retrieval System

Andri Syahrian Eka Putra[1], Allen Adiguna[2], Aang Arif Munandar[3], Mercuris Broto Legowo[4]

[1234]Faculty of Information Technology, Perbanas Institute, Jakarta Indonesia

Jl. Perbanas, Karet Kuningan, Jakarta 12940

Abstract - **The problem of finding student data that is large enough so that it is difficult to retrieve information about students is a problem that occurred at FTI Perbanas Institute Jakarta. Information retrieval is based on queries by users, who are expected to find documents based on user needs to be known as Information Retrieval. One method that can be used for this problem is the VSM or Vector Space Model method. The purpose of this study is to implement the Vector Space Model for search engines to be able to provide text document information in the right database using certain keywords. The results of the search are represented by the order and ranking of similarities of documents with queries. The expected results are a system application that makes it easier for users to be able to collect data or documents that are owned quickly, easily and accurately. This system application will be a very useful tool for optimal data search.**

Keywords - Information retrieval, queries, vector space model, document, system.

I. INTRODUCTION

Information retrieval (IR) system is a system used to retrieve information that is relevant to the needs of users of a collection of information automatically [1]. Information Retrieval is a part of computer science that deals with retrieving information from documents based on the content and context of the documents themselves. The process of Information Retrieval can be described as a process to obtain relevant documents from collection documents through query searches entered by the user [2].

The Faculty of Information Technology Perbanas Institute Jakarta is one of the campuses that provides various information documents in digital form. The document continues to grow all the time so that the document gets more and more. To find these documents, it takes a relatively long time because the search is still done manually. Therefore we need a data search application that can search documents more quickly, easily and produce relevant information. In this data search application, the author will apply the Vector Space Model method. The method is a method to see the level of closeness or similarity term by weighting the term. Documents are viewed as a vector that has a magnitude (distance) and 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 [3].

Research on Information Retrieval has also been conducted by Hendra Bunyamin and Puspa Negara (2008) who conducted a study on "CATA Information Retrieval (IR) Application with the Generalized Vector Space Model" [1]. the study talked about building a desktop-based data search application using the Vector Space Model method. In contrast to the research conducted by Harry Wardana, Yanuar Firdaus, and Warih Maharani (2010) who conducted a study on "Re-Updating the Index with the B-Tree Method on Information-Based Desktop Search" [4]. This study describes the desktop-based Information Retrieval System with the B-Tree method.

In this research aims to create a data search application that can help users, especially the Faculty of Information Technology Perbanas Institute Jakarta in searching for information documents to be faster, easier and more accurate by applying the Vector Space Model method.

II. LITERATURE REVIEW

This application is a subclass of computer software that utilizes the ability of the computer directly to perform a task that the user wants [5]. the implementation of the Vector

Corresponding Author: Andri Syahrian Eka Putra

Space Model method on system applications is an idea that is quite helpful in data search.

a. *Vector Space Model*

Vector Space Model (VSM) is a method to see the level of proximity or similarity term by weighting the term [6]. Documents are viewed as a vector that has distance and direction. The relevance of a document to a query is based on similarity between document vectors and query vectors. In the Information Retrieval system, the similarity between documents is defined based on bag of words representation and converted to a vector space model. Dimensions according to the number of terms in the document involved. In this model:

Documents and queries are expressed as dimensional t vector dj = (w1, w2, ..., wtj) and there are n documents in the collection, namely j = 1, 2, ..., n. An example of a three-dimensional vector space model for two documents D1 and D2, one user query Q1, and three terms T1, T2 and T3.

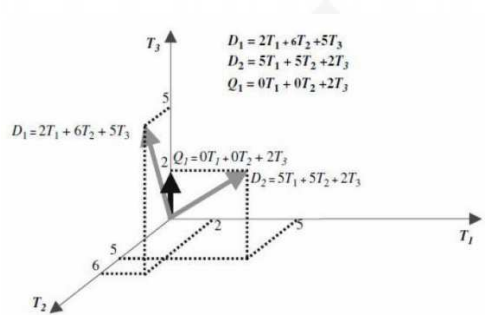

Figure 1. Economic Order Quantity [7].

b. *Indexing*

In the domain information retrieval, the indexing process is a very important process. The indexing stage is done to save each word into the database along with the weight of each word. By using indexing the information seeker can easily find the information he wants [7].

c. *Parsing*

Parsing is the process of introducing tokens (tokenizing) in a text string. In making a term index, documents are broken down into smaller units such as words, phrases or sentences. These units are usually referred to as tokens [8].

d. *Precision & Recall*

Precision is one quantity that is commonly used to measure the performance of an information retrieval engine [9]. At the same time, recall is a number of relevant documents that were found in the search results [9].

e. *Previous Related Research*

There are several previous studies related to Information Retrieval.

Research conducted by Harry wardhana, Yanuar Firdaus, and Warih Maharani explain the desktop-based Information Retrieval System with the B-Tree method, where this method is one method in indexing which generally develops from the bottom up in accordance with the elements that are filled [4].

Other research has also been carried out by Hendra Bunyamin and Puspa Negara, who talked about building desktop-based applications that can make data search more effective and efficient using the Vector Space Model method [1].

From the review of the study above, it can be concluded that there are differences in research that will be made with previous research. The difference can be seen from the process of making the application, if in the previous research the process of making an application was only offline or only a desktop-based application, then this research will be online and website-based.

f. *Theoretical Framework*

The mindset of the Information Retrieval is divided into two parts. The first part is part of the information seeker or user of the system. Users of the Information Retrieval system must translate the information they are looking for so that it can be processed by the system by entering keywords. These keywords will be processed into a query that can be understood by the computer. The second part is part of the document. In this section the documents are represented in the form of an index. Later queries from users will be processed through a similarity function to compare queries with indexes of documents to obtain relevant documents [8].

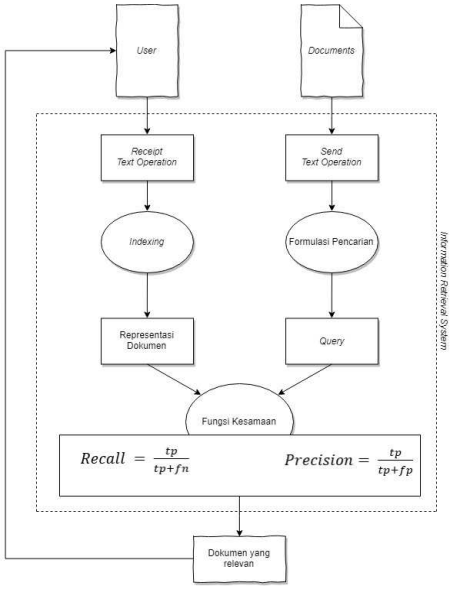

Figure 2. Theoretical Framework

III. METHODOLOGY

This type of research is an applied product research definition of research that is directed towards creating innovation and technological development by using a qualitative descriptive approach. In this study using all FTI Perbanas Institute Jakarta Student data.

In this research, the stages starting from data collection obtained from reference studies as well as preliminary observations at the research location, the second stage in the form of initial data processing, carried out by understanding the existing system and knowing the shortcomings of the running system, third stage in the proposed method or model by making a conceptual framework as a system design, the fourth stage is, testing the model or system and evaluating using the black box and white box methods, the fifth stage of the model or system validation stage, using this person or expert or commonly called the expert confirmation test.

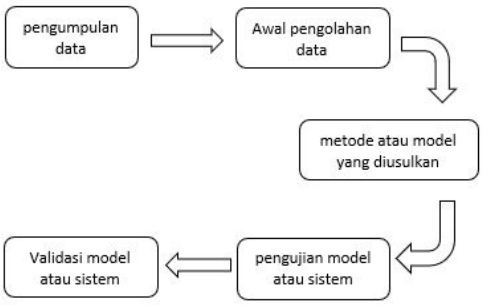

Figure 3. Stage of Research

In this study using the SDLC method or the System Development Life Cycle (SDLC).

The SDLC method is a methodology used to develop, maintain, and use information systems [10]. There are various models of software development processes, including the Linear Sequential Model. This model consists of several stages, namely: Analysis (system requirements analysis), Design (design), Coding (implementation), Testing (testing) [11].

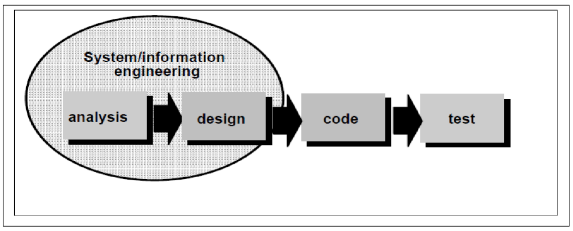

Figure 4. Stages of the SDLC Method [11]

a. *Conceptual Definitions*

The conceptual definition is the meaning of the concept used, making it easier for researchers to operate the concept in the field. To understand and facilitate interpreting many of the theories in this study, several conceptual definitions relating to be studied [12], among others:

1. *Information Retrieval*

By looking at a number of definitions of Information Retrieval by some experts who have been explained in the previous sub-section, Information Retrieval is the science of collecting data in a database, more precisely Information Retrieval is useful for making data searches in the database more effective and efficient. Information Retrieval discussed in this thesis is a document search application that is used as a tool to search for Perbanas Institute's Information Technology Faculty documents that are constantly increasing every day. With this data search application, you can search documents more quickly, easily and produce relevant information [1].

a. *Indexing*

The indexing stage is done to save each word into the database along with the weight of each word. By using indexing the information seeker can easily find the information he wants [7].

b. *Parsing*

Parsing is the process of introducing tokens (tokenizing) in a text string. In making a term index, documents are broken down into smaller units such as words, phrases or sentences. These units are usually referred to as tokens. While the algorithm for solving a collection of sentences or phrases into tokens is called a tokenizer. Parsing processing is strongly influenced by the knowledge of the language used in a document to handle special characters, and to provide boundaries for tokens in a document [8].

c. *Weighting The Term*

The term in an index must be able to distinguish the interests of a document on an information. The method is by giving weight to a term for a document. The higher the weight of a term, the more important the term is compared to other terms in a document. The weight of this term is included in the inverted index to be used in the process of finding documents [8].

d. *Precision & Recall*

To measure the performance of a retrieval engine, at least 2 parameters are used, namely Precision and Recall. Both are quantities that can show the performance of a retrieval engine. The higher the value of both, the better the performance of the retrieval engine [9].

2. *Vector Space Model Method*

Documents are viewed as a vector that has a magnitude of distance and direction of direction. The Vector Space Model (VSM) method is used as a representation of a collection of text document datasets. The importance of a word in the document can be seen from the frequency of its appearance on the document [6]. To distinguish the interests of a document on information. The method is by giving weight to a term for a document. This weight states the

importance or contribution of words to a document and document.

3. *The Research Instrumen*

The research instrument is in the form of a test that is measuring, because it contains questions and statements whose alternative answers have a certain standard of answer, either incorrect or scale of answers [13]. The interview instrument here is a list of questions that will be asked to the parties concerned. For documentation, researchers investigated released objects such as books, archives and documents that will be used as research data. The following questions will be asked at the interview.

1. How much is the FTI Perbanas Institute Jakarta Student data?

2. What problems are commonly found when collecting data from the FTI Perbanas Institute Jakarta Student data?

3. How important is the Student data search application for FTI Perbanas Institute Jakarta?

b. *Data Collection Techniques*

The method used by researchers in analyzing the data is several stages, namely:

1. *Field Study*

a. *Observation*

Observation is an observation activity to find out the problems and the proposed system development survey.

b. *Interview*

This activity was carried out in knowing the need for the system to be built with information obtained from prospective users of student data search applications at Perbanas Institute.

2. *Literature Study*

This section is carried out to obtain written data data both theoretically and empirically related to the research topic. Besides the literature study conducted by researchers, based on the collection of materials related to the title of research proposals through reading books and internet sites, research related to the creation of a website-based Information Retrieval System.

c. *Types & Data Sources*

The data source is where data is obtained using certain methods in the form of humans, artifacts, or documents [14]. The following types of data according to the source.

1. *Primary Data Sources*

Primary Sources is the data obtained directly includes company documents in the form of the history of company development, organizational structure and others related to research [15]. The primary data source in this study is one of the lecturers at FTI Perbanas Institute. Through interviews, information was obtained directly about FTI Perbanas Institute student data.

2. *Secondary Data Sources*

The Source of Secondary Data is the data needed to support the results of the study derived from literature, articles and various other sources related to research. Secondary data, among others, are presented in the form of data, documents, tables on research topics [15]. Secondary data sources obtained were in the form of processed FTI Perbanas Institute Student data.

d. *Analysis Technique*

According to Taylor (1975) in Stela (2018) it defines data analysis as a process that formally attempts to find themes and formulate hypotheses (ideas) as suggested and as an attempt to provide assistance and themes to the hypothesis[15].

1. *Calculation techniques*

Engineering is a way, tool, or media used by teachers to direct the activities of students towards the goals to be achieved [15]. And the calculation according to the Indonesian General Dictionary written by Poerwadarminta calculation is defined as follows, "Calculation is the sum / determination of total expenditure or payment for a service and delivery costs." From the two opinions above, it can be concluded that the calculation technique is a media or tool used to determine a get the desired results.

a. *Precision*

Precision= tp/(tp+fp)

b. *Recall*

Recall = tp/(tp+fn)

2. *System Test*

System testing is used as the main part of the development process, why? Because in these tests can be known deficiencies while testing whether the system is in accordance with needs. There are several ways commonly used in testing the system.

IV. RESULT AND DISCUSSION

On white box testing or structural testing, the test conditions were designed by testing the flow of logic. Testers test the internal structure of the program or system. Data testing is based on testing the logic of the program or system, regardless of program or system requirements. Examiners know the program's internal structure and.

Black box testing is testing software in terms of functional specifications without testing the design and program code. Testing is intended to find out whether the functions, input, and output of the software are in accordance with the required specifications. Black box testing is done by making a test case that has the character of trying all functions using software whether it is in accordance with the required specifications. Test cases made to carry out black box testing must be made with the correct case and wrong case. Later the testing method used by the author is Black-Box testing [16].

V. Conclusion And Recommendation

a. *Conclusion*

Based on the results of analysis, design, implementation and testing of data search applications using the Vector Space Model method to search data at the Perbanas Faculty of Information Technology, the conclusions obtained are as follows;

1. Produce a data search application as an applied product that is very useful for the Perbanas Information Technology Faculty.
2. Produce data search applications that have calculated index similarities with VSM and provide accurate results.

b. *Recommendation*

In order to develop further data search applications in the future, the authors suggest that this system application needs to be added to a module related to student data for the data storage process that is integrated with this application.

Acknowledgement

The author would like to express his gratitude to the Directorate of Research and Community Service of Perbanas Institute - Jakarta, for their assistance in giving this Research Fund in 2018.

Finally, the author also expresses the deepest gratitude to the Dean of the Faculty of Information Technology for providing support and opportunities to conduct research related to study programs, and implement this system in the Computer Science Study Program and other study programs at Perbanas Institute in the future.

References

[1] H. Bunyamin and C. P. Negara, "Aplikasi Information Retrieval (IR) CATA Dengan Metode Generalized Vector Space Model," J. Inform., vol. 4, no. 1, pp. 29–38, 2008.

[2] N. A. Suwastika, Y. F. A, and D. S. Kusumo, "ANALISIS DAN IMPLEMENTASI VECTOR SPACE MODEL (VSM) PADA INFORMATION RETRIEVAL," Tugas Akhir, 2009.

[3] F. Amin, "IMPLEMENTASI SEARCH ENGINE (MESIN PENCARI) MENGGUNAKAN METODE VECTOR SPACE MODEL Fatkhul Amin Dosen Fakultas Teknologi Informasi Universitas Stikubank Semarang," Din. Tek., vol. V, no. 1, pp. 45–58, 2011.

[4] H. Wardana, Y. F. A, and W. Maharani, "Re-updating index dengan metode b-tree pada desktop search berbasis information retrieval," Tugas Akhir, 2010.

[5] N. Safaat, "Pemrograman Aplikasi Mobile Smartphone dan Tablet PC Berbasis Android," Book, 2012.

[6] R. Baeza-yates and B. Ribeiro-Neto, "Modern Information Retrieval," Article, no. July 1999, 2015.

[7] P. E. Mas'udia, M. D. Atmadja, and L. D. Mustafa, "Information Retrieval Tugas Akhir Dan Perhitungan Kemiripan," J. SIMETRIS, vol. 8, no. 1, pp. 355–362, 2017.

[8] F. Nadirman, "Sistem Temu-Kembali Informasi Dengan Metode Vector Space Model Skripsi Sistem Temu-Kembali Informasi Dengan Metode Vector Space Model," Skripsi, 2006.

[9] W. Hidayat, "Indexing and Retrieval Engine untuk Dokumen Berbahasa Indonesia dengan Menggunakan Inverted Index," Semin. Nas. Inform. dan Apl. 2015, no. October, 2013.

[10] J. O'Brien, "Introduction to information systems: Essentials for the internet worked e-business enterprise," 10th Ed. IrwinMcGrawHill, New York, 2001.

[11] S. P. Roger, "Rekayasa Perangkat Lunak Pendekatan Praktisi (BUKU SATU)," ANDI. Yogyakarta, 2002.

[12] M. Singarimbun, "Tipe, Metode dan Proses Penelitian" dalam Masri Singarimbun & Sofian Effendi (ed) Metode Penelitian Survei," Jakarta: LP3ES, pp. 3–24, 1988.

[13] N. S. Sukmadinata, "Metode penelitian pendidikan," Progr. Pascasarj. Univ. Pendidik. Indones. dengan PT Remaja Rosdakarya, 2005.

[14] H. B. Sutopo, "Pengantar Penelitian Kualitatif," Surakarta UNS, 2006.

[15] S. Veranita and M. Broto, "Penerapan Metode Economic Order Quantity Pada Rancangan Aplikasi Inventory Control System," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 2, no. 3, pp. 768–774, 2018.

[16] A. S. Rosa and M. Shalahudin, "Pembelajaran Rekayasa Perangkat Lunak," Bandung. Modul., 2011.