Document Search in Information Retrieval System Using Vector Space Model

2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE) | 978-1-6654-3232-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICEEIE52663.2021.9616735

Yusrandi
Faculty of Engineering
Universitas Negeri Malang
Malang, Indonesia
yusrandhyr@gmail.com

Muladi
Faculty of Engineering
Universitas Negeri Malang
Malang, Indonesia
muladi@um.ac.id

Harits Ar Rosyid
Faculty of Engineering
Universitas Negeri Malang
Malang, Indonesia
harits.ar.ft@um.ac.id

Abd Kadir Mahamad
Faculty of Electrical and Electronic Engineering
Universiti tun Hussein onn Malaysia
parit raja, Johor, Malaysia
kadir@uthm.edu.my

***Abstract*— The current pandemic has spread everywhere. Various effects of pressure in the economic, educational, and social sectors are forced to adjust. So that people really need information about efforts to prevent the spread is very necessary. Search Engine is a program that is used as a tool to find more information on the internet. Search Engine is one of the discussions in the field of Information Retrieval. This system is a document search of unstructured properties. Thus, being able to provide the information needs of a large set of documents (on a local computer server or the internet). The vector space model is one of the many models in Information Retrieval that is used to get the distance and direction between keywords and documents by representing them into vectors. Then the results of ranking using cosine similarity with a dataset of 90 articles about covid19 along with 4 keywords will be tested with precision, recall, and accuracy calculations. The results of the precision calculation get a value of 60% - 73%, recall gets a value for each of the keywords 81% - 100% and gets an accuracy value of 85% - 89%. The results of these experiments indicate that information retrieval with vector space model is effective with good and stable performance used for information retrieval.**

Keywords—Search Engine, request, query, Text mining, Term Weghting, Similarity, Vector Space Model, Precision, recall, Accuracy

I. INTRODUCTION

COVID-19 is classified as a pandemic disease, this is because of the deadly nature of the disease and it is very easy to spread widely [1][2]. The spread of COVID-19 has directly impacted the community and put pressure on it. The most important thing to do right now is to prevent the spread of COVID-19. Then a set of health protocols was issued as an effort by the government to respond to this situation.[3] This effort actually comes from WHO directives, including keep the distance, reducing interaction with others, using masks when doing activities outside the home, washing hands with soap, and keep the environment clean. In order for this prevention to be carried out optimally, socialization to the community is needed. Socialization and education about COVID-19 possible to do through the media or the internet[4]. There are two kinds of people in getting information. Conventional society who seeks news from print media, and modern society that uses the internet to searching for important and actual information[5]. Internet service that has been developed is currently used as a search for information is a search engine.

Google, Yahoo, MSN / Bing and Ask are some of the most popular search engines in the community. Search Engine is an intelligent technology that is very popular and very valuable [6][7]. Search Engine is one of the discussions in the field of Information Retrieval. This system is a document search of unstructured properties. Thus, being able to provide the information needs of a large set of documents (on a local computer server or the internet). information storage and information retrieval is simple. there is a storage area for documents which is often called a corpus and the user formulates a question (request or keyword) whose answer is a set of documents containing the required information expressed through user questions.[5]. Search Engine is one of the discussions in the field of Information Retrieval.

The Information Retrieval System has three models, including: Boolean model, probabilistic model and vector space model [8]. The vector space model is one of the most widely used models for document retrieval, mainly because of its conceptual simplicity and metaphorical appeal that underlies the use of spatial proximity for semantic proximity. The resulting representation is weighted indexing terms. The users do not need to express their information needs by using logical operators. This model provides a more user-friendly access to information[7]. this model will also be used in this study.

II. LITERATURE REVIEW

A. Information Retrieval

Information Retrieval System is a system that finds information according to user needs from a collection of corpus automatically. The way this system works is if there is a collection of documents and a user formulates a question (request or query). The answer to this question is to set the relevant documents and discard the irrelevant documents [5]. Information retrieval is an advanced technique of Natural Language Processing, which is a field of science to increase the effectiveness of word or term based document retrieval [8].

B. Text Mining Process

There are some stages in the text mining process:

a. Parsing

This process is the first stage before term weighting. Tokenization is a language-dependent approach, including normalization, stopword

removal, andstemming. The steps taken are to break a sentence into words. This is done because the weighting at the next stage is carried out in words, not sentences.

b. *Stoplist*

Unimportant words will be discarded. Usually unimportant words include: whitespace characters such as enter, tabulation, and spaces. Affixes such as "read". Conjunctions such as “to”, “at”, “which”, “is”, and punctuation marks such as periods (.) commas.(,) and so on. Removal stopword serves to reduce index and processing time. Can also reduce noise level[9].

c. *Stemming*

Term Operation is the main thing in the IR system, including stemming operations, namely the process ofremoving/cutting from a word into its basic form, cutting, weighting, and eliminating stoplists, namely the process of removing words such as: but, that is, while, and so on.[10]. Stemming is done to reduce the number of different indexes of a document. The stemming technique also works to group other words that have the same root word and meaning but form, which is different because of different affixes[11].

C. Term Weighting (Tf-Idf)

Term Frequency-Inverse Document Frequency (TF-IDF) is a method of assigning weight to the relationship of a word (term) to a document. For a single document each sentence as a document. This method combines two concepts for weight calculation, There are Term Frequency (TF) is the frequency of occurrence of words in sentences. Document Frequency (DF) is the number of times the word is present in all documents[12].

The formula for calculating the Tf-Idf is the following:

$$TF - IDF\ (t_k\ , d_j) = TF(t_k, d_j) * IDF(t_k)$$

Where t_k is a ke-k term and d_j is ke-j document. Where previously calculated the term Frequency (TF) is the frequency of occurrence of a term in each document. Then the Inverse Document Frequency (IDF) is calculated, which is the weight value of a term calculated from the number of times a term appears in several documents. The more a term appears in many documents, the smaller the IDF value will be.

TF and IDF equation:

$$TF(t_k, d_j) = f(t_k, d_j)$$

with TF is sum of term frequency, f is the number of output frequencies, d_j is ke-j document, and t_k is ke-k term. Then formula for calculating the IDF is the following:

$$\mathrm{IDF(t)} = \frac{1}{df(t)} \text{ or } \mathrm{IDF(t)} = \log(\mathrm{N/df})$$

Where $IDF(t_k)$ is term weight, N is sum of document, df is document frequency, d_j is ke-j document and t_k is ke-k term[12].

D. Vector Space Model

The vector space model represents a document in terms of word frequency as a vector in a high-dimensional vector space. This model provides a partial match framework, this is achieved by non-binary weighting for indexing Terms in keywords and documents. This method looks at the level of closeness or similarity (similarity) of the term with the weighting of the term. Documents are represented as vectors that have magnitude (distance) and direction (direction). In the Vector Space Model, a term is represented by the dimensions of the vector space. The relevance of the document to keywords is based on the similarity between the document vector and the keyword vector. The vector is calculated using the cosine similarity function[13].

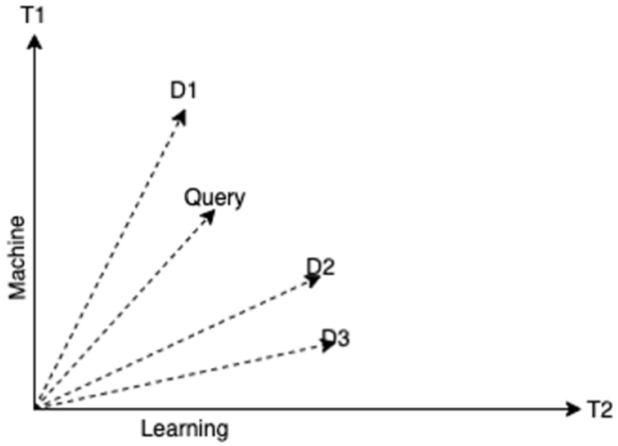

Fig. 1. Vector Space Model Diagram

Documents and queries are represented in a multi-dimensional space. Where each dimension of space, corresponds to the word in the document. The most relevant documents for the query are those represented by the vector closest to the query, i.e. documents that use words that are similar to the query.[14]. According to the Fig. 1. below

The information received by the term frequency is how important a word is in a particular document. The higher the term frequency (the more often the word occurs) the more likely it is that the word is a good description of the document.

For this weight using formula is the following stage :

- Calculate distance of document

$$|d_j| = \sqrt{\sum_{j=1}^{t} (W_{ij})^2}$$

with $|dj|$ is a document distance, and Wij is a weight ke-I documet then the distance of the document is calculated to get the distance of the document from the weight of the document (Wij) taken by the system. Document distance can be calculated by the equation of the square root of the document.

- Calculation of similarity between query and document

$$sim(q, d) = \sum_{i=1}^{t} \mathrm{W_{iq}} \,.\, \mathrm{W_{ij}}$$

where $sim(q, d_j)$ is similarity between *query* and document, W_{iq} is *query weight*, and W_{ij} is a term weight in document. Similarity between query and document is used to get the weight based on the weight of the term in the document

and the query weight by adding up the query weight multiplied by the document weight.[15].

III. METHOD

A. System Architecture

a. Information Retrieval System

In Fig. 2 there are two main processes, pre-processing the database and then applying certain methods to calculate the proximity (relevance or similarity) between documents in the database that have been preprocessed with user queries. the user-entered keywords are converted according to certain rules to extract important Terms that are in line with the previously extracted Terms from the document and calculate the relevance between the keywords and the document based on those Terms. As a result, the system returns a list of documents sorted according to their similarity value to the user's keyword. Terms are stored in a special search database organized as an inverted index. This index is a conversion of the original document containing a set of words into a list of words associated with the related document in which the words appear. The process in the Information Retrieval System can be described as a process to obtain relevant documents from an existing collection of documents through the search for keywords entered by the user.[16].

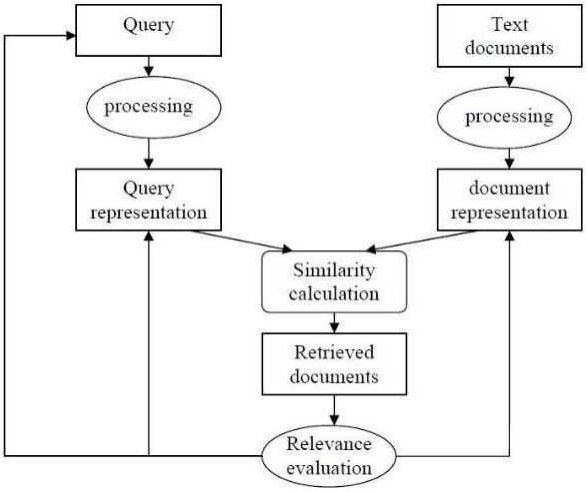

Fig. 2. Information Retrieval Architecture

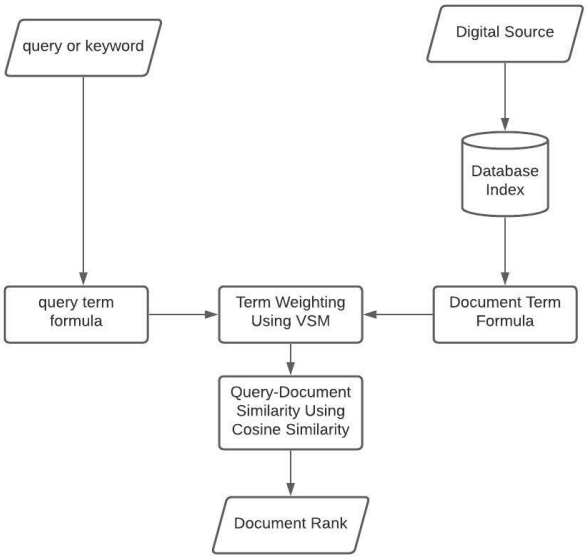

Fig. 3. Vector Space Model Architecture

b. Vector Space Model

Fig. 3 shows that there are three operating steps in an Information Retrieval System using a vector space model.

The first step starts from the collection of documents in the form of digital sources (can be seen in the arrows) to the process of forming an index database. The second step starts from the document search keywords by the user. In these keywords, keyword terms will be formulated, namely calculating the weights of these keyword terms usingthe TF-IDF weighting algorithm then calculate similarity of query and document. While the third step is the document ranking process using the vector space model algorithm[14].

In the vector space model, keywords and documents have weight vectors for words (terms). The similarity of documents and keywords is calculated based on the distance between the document vector (D) and the keyword (Q) using the inner product formula (dot product). The distance is calculated by measuring the angle cosine[13].

B. Dataset

First stage in this research is data collection. The data used is COVID-19 data. This study uses 90 data in the form of articles from the latest information website for handling COVID-19 in Indonesia https://covid19.go.id/edukasi/materi-edukasi

IV. RESULT AND DISCUSSION

In this study, the researcher used the TF-Idf method to get the term weighting, then the vector space model calculated the distance and direction between keywords and documents, then in the ranking stage using Cosine Similarity.

In the testing phase, the researcher used 4 keywords in 90 article documents that were inputted randomly and returned the documents according to each keyword.

TABLE 1. KEYWORD LIST

	Keyword
KK1	Pembelajaran tatap muka di sekolah
KK2	Ibu menyusui dan anak
KK3	Vaksin dan Vaksinasi
KK4	Tenaga Kesehatan

From the results of ranking documents using the keywords in Table 1. Can measure the effectiveness of the system and assess the quality of text retrieval by looking at the value of the calculation of accuracy (precision), the value of recovery (recall) and good accuracy[19] .

And the precision test is the ratio of the relevant documents that were found from all the documents found with formula :

$$precision : \frac{TP}{TP + FP} x\ 100\%$$

Recall testing is the ratio between relevant documents that have been found from all relevant documents in the system. With formula :

$$recall : \frac{TP}{TP + FN} x\ 100\%$$

Then the Accuracy test is the calculation of the ratio between the correct predictions from positive or negative documents with the overall data in the collection. With formula below :

$$accuracy : \frac{TP + TN}{TP + TN + FP + FN} \; x \; 100\%$$

TABLE 2. KK1 CALCULATION RESULT

Pembelajaran tatap muka di sekolah		True Value	
		TRUE	FALSE
Predicted Value	TRUE	22	8
	FALSE	5	57

$$precision : \frac{22}{22 + 8} \; x \; 100\% : 0{,}73 \; x \; 100\% : 73\%$$

$$recall : \frac{22}{22 + 5} \; x \; 100\% : 0{,}81 \; x \; 100\% : 81\%$$

$$Accuracy : \frac{22 + 57}{22 + 57 + 8 + 5} \; x \; 100\% : 0{,}85 \; x \; 100\% : 85\%$$

TABLE 3. KK2 CALCULATION RESULT

Ibu menyusui dan anak		True Value	
		true	false
Predicted Value	true	21	8
	false	3	58

$$precision : \frac{21}{21 + 8} \; x \; 100\% : 0{,}74 \; x \; 100\% : 74\%$$

$$recall : \frac{21}{21 + 3} \; x \; 100\% : 0{,}87 \; x \; 100\% : 87\%$$

$$Accuracy : \frac{21 + 58}{21 + 58 + 8 + 3} \; x \; 100\% : 0{,}87 \; x \; 100\% : 87\%$$

TABLE 4. KK3 CALCULATION RESULT

Vaksin dan Vaksinasi		True Value	
		TRUE	FALSE
Predicted Value	TRUE	22	10
	FALSE	0	60

$$precision : \frac{22}{22 + 10} \; x \; 100\% : 0{,}68 \; x \; 100\% : 68\%$$

$$recall : \frac{22}{22 + 0} \; x \; 100\% : 1 \; x \; 100\% : 100\%$$

$$Accuracy : \frac{22 + 60}{22 + 60 + 10 + 0} \; x \; 100\% : 0{,}89 \; x \; 100\% : 89\%$$

TABLE 5. KK4 CALCULATION RESULT

Tenaga Kesehatan		True Value	
		TRUE	FALSE
Predicted Value	TRUE	15	10
	FALSE	1	66

$$precision : \frac{15}{15 + 10} \; x \; 100\% : 0{,}60 \; x \; 100\% : 60\%$$

$$recall : \frac{15}{15 + 1} \; x \; 100\% : 0{,}93 \; x \; 100\% : 93\%$$

$$Accuracy : \frac{15 + 66}{15 + 66 + 10 + 1} \; x \; 100\% : 0{,}88 \; x \; 100\% : 88\%$$

TABLE 6. OVERALL RESULT OF KEYWORD CALCULATION

Keyword	*precision*	*recall*	*accuracy*
Pembelajaran tatap muka di sekolah	73%	81%	85%
Ibu menyusui dan anak	72%	87%	87%
Vaksin dan Vaksinasi	68%	100%	89%
Tenaga Kesehatan	60%	93%	88%

V. CONCLUSIONS

The search for information in the vector space model is based on the similarity between keywords and documents. Longer documents with a higher number of Terms are more likely to be considered relevant to certain keywords than shorter documents.

Based on the data from the recall test results, the value of 81% - 100% indicates that the system with the vector space model has succeeded in getting all documents relevant to the keywords desired by the user. and precision gets a value of 60% - 73% which indicates the system finds other documents that are not relevant to the keywords, because the precision value depends on the uniqueness of the keywords entered by the user. And get an accuracy value of 85% - 89%.

For further research, we can optimize the weighting of the vector space model or can do query expansion to get more and more relevant documents.

REFERENCES

[1] S. Khan *et al.*, "The spread of novel coronavirus has created an alarming situation worldwide.," *J. Infect. Public Health*, vol. 13, no. 4, pp. 469–471, Apr. 2020, doi: 10.1016/j.jiph.2020.03.005.

[2] D.-G. Ahn *et al.*, "Current Status of Epidemiology, Diagnosis, Therapeutics, and Vaccines for Novel Coronavirus Disease 2019 (COVID-19).," *J. Microbiol. Biotechnol.*, vol. 30, no. 3, pp. 313– 324, Mar. 2020, doi: 10.4014/jmb.2003.03011.

[3] T. C.-19 I. Erlina Burhan, Agus Dwi Susanto, Sally A Nasution, Eka Ginanjar, Ceva Wicaksono Pitoyo, Adityo Susilo, Isman Firdaus, Anwar Santoso, Dafsah Arifa Juzar, Syafri Kamsul Arif, Navy G.H Lolong Wulung, Triya Damayanti, Wiwien Heru Wiyono, Prasenohadi, Afiatin, "Protokol Tatalaksana Covid-19," *1*, pp. 1–50, 2020.

[4] S. A. Prayitno, H. P. Pribadi, and R. A. Ifadah, "Peran Serta Dalam Melaksanakan Protokol Pencegahan Penyebaran Corona VirusDisease (Covid-19) Pada Masyarakat," *DedikasiMU (Journal Community Serv.*, vol. 2, no. 3, pp. 504–510, 2020.

[5] I. Irmawati, "Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model," *J. Ilm. FIFO*, vol. 9, no. 1,p. 74, 2017, doi: 10.22441/fifo.v9i1.1444.

[6] N. Vincent, I. Johnson, P. Sheehan, and B. Hecht, "Measuring the importance of user-generated content to search engines," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2019, vol. 13, pp. 505–516.

[7] R. Febriyan, "Model Ruang Vektor pada Information Retrieval System," pp. 2–5, 2016.

[8] B. A. Abu-Salih, "Applying Vector Space Model (VSM)Techniques in Information Retrieval for Arabic Language," *arXiv*,2018.

[9] A. Hendrawan, T. Winarti, and H. Indriyawati, "Pengembangan stemming untuk artikel berbahasa indonesia," 2020.

[10] K. D. Putung, A. S. M. Lumenta, and A. Jacobus, "Penerapan Sistem Temu Kembali Informasi Pada Kumpulan Dokumen Skripsi," *J. Tek. Inform.*, vol. 8, no. 1, 2016, doi: 10.35793/jti.8.1.2016.12227.

[11] A. Samir and Z. Lahbib, *Stemming and lemmatization for information retrieval systems in amazigh language*, vol. 872. Springer International Publishing, 2018.

[12] B. Poernomo *et al.*, "Sistem Information Retrieval Pencarian Kesamaan Ayat Terjemahan Al Quran Berbahasa Indonesia," *Semin. Nas. Teknol. Inf. dan Komun.*, pp. 100–108, 2015.

[13] S. Pencarian, I. Pemesanan, M. Online, and M. Vektor, "JurnalMantik," vol. 5, no. 36, pp. 93–97, 2021.

[14] L. B. Doyle, "Information retrieval," *Commun. ACM*, vol. 4, no. 4, p. 195, 1961, doi: 10.1145/355578.366528.

[15] E. Wahyudi, "Sistem Pencarian Informasi untuk Pencarian JSON File dengan Metode Model Ruang Vektor," pp. 260–265, 2019.

[16] N. T. Wai Khin and N. N. Yee, "Query Classification based Information Retrieval System," *2018 Int. Conf. Intell. Informatics Biomed. Sci. ICIIBMS 2018*, vol. 3, pp. 151–156, 2018, doi: 10.1109/ICIIBMS.2018.8549988.

[17] N. E. Rozanda, A. and Marsal, and K. Iswanti, "Rancang Bangun Sistem Informasi Hadits Menggunakan Teknik Temu Kembali Informasi Model Ruang Vektor," p. 8, 2012.

[18] A. Hadhiatma, "Pencarian Dokumen Berdasarkan Kombinasi Antara Model Ruang Vektor dan Model Domain Ontologi," *Semin. Nas. Inform.*, vol. 1, no. 4, pp. 111–117, 2010, [Online]. Available: http://jurnal.upnyk.ac.id/index.php/semnasif/article/view/1189.

[19] S. Bahri, "Aplikasi Pencarian Bahan Pustaka Di Perpustakaan Menggunakan Metode Vector Space Model," *J I M P - J. Inform. Merdeka Pasuruan*, vol. 5, no. 2, 2021, doi:10.37438/jimp.v5i2.265.