Literatify : Trends in Library Developments
2722-1571 (p-ISSN)
2723-0953 (e-ISSN)
Volume 5 Issue 1, March 2024

A Literature Review: the Importance of Term Normalization in Vector Space Model

Fayyaz Mubarak Hasyim[1] & Faisal Fahmi[2]

[1,2]Departement of Information and Library Science, Airlangga University
Correspondence Email: faisalfahmi@fisip.unair.ac.id

Abstrak

In today's digitally inundated era, accessing information is more accessible yet challenging due to the sheer volume available. This article underscores the pivotal role of VSM in managing vast data and enhancing retrieval accuracy by ranking documents based on query similarity. Term normalization, a part of VSM development, standardizes words for indexing, improving accuracy by addressing word variations. The study's methodology involved a systematic literature review, data collection via electronic databases, and thematic analysis. The research findings highlight vital aspects: the fundamentals of information retrieval systems, the working principle of VSM in document sorting, and the process of term normalization. Various methods within term normalization, such as tokenizing, filtering, stemming, and term weighting (e.g., TF, IDF, Cosine Similarity), are elucidated for refining document relevance. Discussions underscore the impact of term normalization on information retrieval, emphasizing heightened accuracy, efficiency, and reduced error rates. In the research paper, five studies that showcased successful applications of VSM across diverse domains were referenced. These domains included karaoke song searches, thesis examiner selection, pest identification in rice plants, hadith interpretation, and library material searches. Each study demonstrated the effectiveness and versatility of VSM in solving various problems in different fields. In conclusion, VSM emerges as a potent tool in managing information overload, particularly when coupled with normalization techniques. The studies reviewed illustrate VSM's efficacy in delivering precise results, affirming its status as a preferred method in information retrieval systems due to its accuracy and effectiveness.

Article Info

Sumitted: 07-01-2024
Review: 04-03-2024
Accepted: 26-03-2024

DOI:
10.24252/literatify.v5i1.44458

How to Cite: Hasyim, F. M., & Faisal Fahmi. (2024). A Literature Review: the Importance of Term Normalization in Vector Space Model. Literatify : Trends in Library Developments, 5(1). https://doi.org/10.24252/literatify.v5i1.44458

Keywords: Vector Space Model (VSM); Information Retrieval; Term Normalization; Information Technology

A. Introduction

In the era of massively moving digital information, accessibility to obtain information is easy and has many variations compared to before. This phenomenon directly illustrates the rapid development of information technology, information, and more and more digital sources. As a result of these developments, the bar's information capacity inevitably continues to increase, leading to unlimited access to knowledge. However, this also raises considerable obstacles. With so much

information accessible, information retrieval becomes a complex environment that can pressure users. To find the information they need, users have to put in extra effort to find information relevant to their needs. Users are confused and surprised by the various results of information searches obtained through search engines. When people receive less information, they experience delays in finding information that suits their needs and targeted topics. This can slow down the information retrieval process.

Information retrieval is one of the most suitable concepts for dealing with such obstacles. Information retrieval systems are a pillar in digital information with a critical role in understanding and managing large amounts of information. The success of these systems dramatically impacts the user's experience and ability to find information relevant to their needs. This has made information retrieval systems more critical than ever. In addition, the role of information retrieval systems as presenters of accurate search results has become very significant among users.

The importance of information retrieval systems revolves around understanding the context in which digital information is being developed and improving methods and tools to provide maximum results. In information retrieval, several approaches can help information seekers find relevant information. Vector Space Model (VSM) can be used as one of the ways to overcome this phenomenon. VSM can help reduce the number of relevant documents by ranking documents based on their similarity to the query. Documents most similar to a query or keyword will be marked with a higher rank, allowing users to focus on the most relevant documents. VSM can also improve the accuracy of information retrieval by considering the similarity between query and document. In addition, this model approach can help users to get the most suitable documents that match the given query or keywords. Thus, VSM can help overcome such problems by reducing the number of relevant documents, improving the accuracy of information retrieval, and increasing the efficiency of information retrieval. With such capabilities, VSM has become one of the mainstays in document clustering based on the query given by the user. Although VSM has become the basis of many information retrieval applications, there are still complex obstacles in managing digital documents that require a more sophisticated approach to VSM itself.

To improve the VSM method and the user experience in finding relevant information, experts created the term normalization in response to these complex issues. Term normalization is a process in which the words in a document are transformed into a standardized form so that information retrieval systems can index and quickly process them. In addition, normalization also plays a role in identifying queries, reducing word ambiguity, and building a vector representation to represent each managed document. For example, the words "child" and "children" should be considered words with the same context even though there are differences in the writing of words. Without normalization, the results generated from the VSM

process can be distorted and reduce the accuracy of the information provided to users.

With so many aspects of term normalization, it is evident that term normalization has become a part of VSM development for the better. As such, this article will discuss the importance of term normalization in VSM to produce relevant information findings for users. Based on the problems stated, this article aims to explain how vital term normalization in VSM is to the results given to information users by showing some results from implementations carried out by other researchers regarding VSM or term normalization. In its discussion, this article covers several areas that are relevant to the topic at hand. Here are some of the discussions:

1. A brief explanation of information retrieval systems.
2. An explanation of VSM as the basis of an information retrieval system.
3. Explanation of term normalization in the context of VSM.
4. Explanation of the impact of term normalization on information retrieval.
5. Explanation of the importance of term normalization in VSM.

B. Research Methodology

This chapter presents the research methodology used to carry out the study titled "A Literature Review: The Importance of Term Normalization in Vector Space Model." The chapter presents the research design, data collection methods, data analysis methods, and ethical considerations. The study adopted a systematic literature review design. The systematic literature review design is a rigorous and transparent method of identifying, evaluating, and synthesizing relevant research studies. The design was appropriate for this study as it allowed for identifying and evaluating relevant literature on the importance of term normalization in vector space models. The data collection process involved using electronic databases such as Google Scholar. The search terms used were "term normalization," "vector space model," "information retrieval," and "text processing." The inclusion criteria for the study were that the articles should be published between 2010 and 2021 and written in Indonesia. The exclusion criteria were that the articles should not be conference papers, they should not be duplicates, and they should not be irrelevant to the study. The data analysis process involved the use of a thematic analysis approach. The thematic analysis approach involved the identification of themes and patterns in the data. The themes were then categorized into sub-themes, and the findings were presented in a narrative format. The study adhered to ethical considerations such as confidentiality, anonymity, and informed consent. The study did not involve human subjects; therefore, there was no need for ethical clearance.

C. Result Findings

Basic Concepts of Information Retrieval Systems

The basic concept of an information retrieval system is that it can provide appropriate and suitable information to information users to meet their needs. As information technology develops, retrieval systems develop in various fields with a strong relationship with information and information users, such as digital libraries, customer relationship management systems, and document collections. In the process of searching for information, information retrieval systems involve the following concepts:

User's query: The user enters a query or keywords in words, phrases, or more complex search forms to the search engine to determine and find the information needed.

Indexing: Documents in the dataset are indexed by the system. The documents are organized by content, keywords, or other attributes that are still relevant to the document. This allows information retrieval to work faster and more efficiently.

Matching: The information retrieval system matches the query given by the user with the indexed documents to identify which document best matches the query entered.

Ranking: After retrieving the documents relevant to the query, the system ranks them based on their relevance to the user. In this process, various factors, such as the document's popularity, the user's preference, or the frequency of the document's query, influence the document's ranking.

Presentation: By presenting the document in a way that is easy for the user to understand and navigate, the system can help users quickly find the information they need, improving their overall experience and satisfaction with the system. This can be in the form of presenting a snapshot of the document, highlighting the query entered by the user on the document, or providing sort options to identify the required information more in sync with the existing needs.

Evaluation: Evaluating the system's efficiency in providing information relevant to user needs and evaluating the use of resources.

Improvement: Improving the system based on the evaluation. This could include expanding the query, improving the indexing and query matching algorithms, or applying advanced techniques to improve the information retrieval process so that the information produced is more relevant to the user's needs.

Vector Space Model (VSM)

Vector Space Model (VSM) is one of the algorithm operations that is often used in the operation of this information retrieval system. With this method, various documents will be indexed and sorted according to the weight of the query in the document. The basic concept of this VSM method is to calculate the distance between documents, which will be sorted based on their proximity to the query entered by the user. In VSM. A term is described as a dimension of the vector space. This will later accumulate the level of similarity of each document in the information retrieval system with the query or keywords entered by the user. The VSM method was first introduced by Salton in 1989.

To calculate the VSM, the user needs to take several preliminary steps. These steps can start with the formation of a query vector, calculation of Term Frequency - Inverse Document Frequency (TF-IDF) weights on document and query words, accumulation of cosine similarity, sorting of cosine values, calculation of Inverse Average Precision (IAP) and Normalized Inverse Average Precision (NIAP), and also evaluation of the results generated from the system output. Although the calculation is quite complicated, the results produced are logically acceptable and are one of the appropriate and effective methods for obtaining relevant information for users.

Term Normalization

Term normalization is one of the steps in implementing VSM before starting the calculation to generate relevant information for users. Normalization exists to overcome the imbalance of term weights between queries and documents. The existence of this normalization aims to improve the results of information retrieval to be more relevant and accurate. In its processing, normalization changes the weight of the terms in the document to be more representative.

Normalization is done by running text processing, which is an effort to change certain characters in the document, such as punctuation marks or capital letters. In addition, in normalization, documents will be carried out in the text mining stage, where documents will be analyzed to find specific patterns and representative words so that they can be compared between documents with one another. When doing text mining, several steps must be taken. Here are some of these steps:

Tokenizing: the truncation of words in the document by using spaces as boundaries. These pieces will be generated as word tokens. In cutting, the words in the document will be modified, such as changing uppercase letters to lowercase or vice versa. After that, it will continue with word parsing, where the words that have been cut will be collected, and the relationship between these words according to their role and function in the document will be considered.

Filtering: Filtering words that have been tokenized by removing words that do not contain value or are considered unimportant, such as affix vocabulary or conjunctions that are widely used in a document with a significant accumulation but do not significantly affect the subject discussed in the document.

Stemming: Modifying words with initial or final affixes by trimming them and taking the basic word to its original form.

Term Normalization Method

After doing text processing and text mining, the next stage of the processed words will undergo a term weighting process. Term weighting is a stage where the weights of the words that have been processed will be calculated for similarity between documents to produce a good document ranking on search engines. Here are some methods that play an important key role in doing word weighting:

Term Frequency (TF): calculates the frequency with which a word is mentioned in documents that will be weighted. The weight will be greater if the word is used extensively in the document. There are several ways to calculate TF. Some of these methods are as follows:

Raw TF: The frequency value of a term is calculated based on the number of words in the document.

Logarithmic TF: To get the TF value, a logarithmic function math calculation is performed with the formula:

$$TF = 1+\log (TF)$$

Pict 1: Logarithmic TF Formula

Binary TF: This calculation will produce a Boolean value of the document. The value will be 0 if a term in the document does not appear and will be one if the term appears. However, in this calculation, the number of terms in the document does not affect the value.

Augmented TF: The value of TF will be calculated by the formula:

$$TF = 0.5 + 0.5 \times TFmax (TF)$$

Pict 2: Augmented TF Formula

TF value is the number of terms in the document, and TFmax value is the number of terms in the same document.

Inverse Document Frequency (IDF): This method measures the number of specific terms used to calculate a term in the document. This IDF calculation aims to reduce the weight of general terms often appearing in documents and increase the weight of more specific terms. In its calculation, IDF uses the following calculation:

$$IDF = Log\frac{D}{Df}$$

Where:
IDF = Inverse document frequency
D = Total Documents
Df = Document Frequency of term
Log = To minimize its influence relative to the weight of the term, it is calculated using the formula: $W = Tf \times IDF$

Where:
W = Weight of document
Tf = Term frequency
IDF = Imverse document frequency

Pict 3: IDF Formula

Thus, IDF helps identify the most essential words in a document that match the query and also helps rank the documents in the search results.

Cosine Similarity: It is a normalization method using the similitude function calculation. The function, in general, is a calculation that accepts two items and returns the similarity weight with a natural number. The resulting value is usually in the interval [0...1]. A value of 1 means that the two objects have similarities, while 0 has no similarities. However, there are still results from the function outside the interval.

In doing the calculation, this method calculates the similarity of two vectors and dimensions by finding the cosine of the angle of the two vectors. The formula for calculating cosine similarity is as follows (Triana, A. 2014)

$$Similarity(x, y) = \cos(\varnothing) = \frac{x.y}{\|x\|\|y\|}$$

Where:

$x.y$: vector dot product of x and y, calculated by: $\sum_{k=1}^{n} x_k y_k$

$\|x\|$: length of vector x, calculated by: $\sum_{k=1}^{n} x_k^2$

$\|y\|$: length of vector y, calculated by: $\sum_{k=1}^{n} y_k^2$

Pict 4: Cosine Similarity Formula

D. Discussion

Impact of Term Normalization on Information Retrieval

Based on how term normalization works, it impacts information retrieval. These impacts are increased accuracy and relevance, increased efficiency, and reduced error rates. Complex calculations generate the information to provide data based on the given query.

To prove this, this article will highlight the results of VSM calculations in previous research:

1. Anna and Ade Hendini wrote this research to discuss the application of the VSM method to the karaoke song machine. In this study, the authors researched the machine's information retrieval system to find karaoke songs by applying VSM calculations as a basis. In conducting their research, the researcher used several samples of song titles to design a karaoke search engine using VSM. After that, the author did the calculation. Based on the study's conclusions, the results of the calculations researchers carried out were successful because they could show which song titles were appropriate and inappropriate for the query entered into the search engine. The researcher also explained that the application of VSM to the song search system based on the title is beneficial in searching for songs synchronized with visitors' needs.

2. The second research was written by Riki Ruli A. Siregar, Fera Amelia Sinaga, and Rakhmat Arianto, discussing making an application for determining thesis examiner lecturers using the TF-IDF and VSM methods. In this study, researchers attempted to create a system application to determine thesis

examiner lecturers as an alternative way for department secretaries to determine thesis examiners. Researchers apply VSM calculations to determine suitable lecturer examiners. The calculation is done by matching the student's thesis title by assessing the similarity of the lecturer's ability in that field. Based on the conclusions made by the researcher, the application of VSM in the system of determining the examiner lecturer successfully matches the existing data with an accuracy rate of 93.22%. So, with this application, Suitable and relevant supervisors can easily find students as the necessary calculations have already been made.

3. The third research was written by Ana Triana, Ristu Saptono, and Meiyanto Eko Sulistyo on using VSM and Cosine Similarity methods to identify pests and diseases in rice plants. In this study, researchers created a system that combines the VSM and Cosine Similarity calculation methods to be used as features to identify pests or diseases that exist in rice plants. The feature will match feedback or input from user answers to detect pests and diseases in rice plants. Based on the study's conclusion, the researcher explained that the calculation of VSM and Cosine Similarity succeeded in identifying the user's input and providing output per the type of pest and disease. In the process, the system has a success rate ranging from 92% to 100%.

4. The fourth research was written by Ria Melita, Victor Amrizal, Hendra Bayu Suseno, and Talismun Dirjam, with a discussion of the application of the TF-IDF and Cosine Similarity methods in the information retrieval system on a hadith matching website. In this study, researchers tried to build a website to find out the syarah of a hadith. Researchers apply the TF-IDF and cosine similarity calculation methods to the system to provide search results. In the process, researchers perform calculations by entering keywords about hadith, such as hadith history and others. After that, the researcher performs the required process. Based on the conclusions written, researchers found that applying the TF-IDF and Cosine Similarity methods was very successful. In addition, remember that in the test, this system gets a value of 100% for precision, 88.7% for recall, 88.73% for accuracy, and 11.27% for error rate. With these revenues, the system is declared excellent and successful by researchers.

5. The fifth research, written by Syaiful Bahri, discussed the creation of library material search services using the VSM method. In the research, researchers tried to create a search engine using the VSM and TF-IDF methods to rank the top 10 documents. Researchers used 150 journals inputted using four selected keywords that had passed the VSM calculation. Based on the conclusion, the researcher explained that the application of the VSM method brought excellent and stable results as it was done in the trial. In addition, the system managed to get a precision of 83% - 100%, recall of 100%, and

accuracy of 96% - 100%, indicating that the system produces relevant information to the query.

E. Conclusion

In overcoming the problem of information overload, the Vector Space Model (VSM) is one of the recommendations to provide information suitable for user's needs and goals. Systems that use VSM as the basis for their work make complex calculations to get the optimal and appropriate calculations. In its implementation, VSM is supported by additional methods such as normalization with the Term Frequency system, Invert Document Frequency, and Cosine Similarity. All three work by pruning the terms in the document and calculating the weight to rank which documents are very similar to the query entered by the user to the system.

The application of the VSM method in information retrieval systems has shown excellent performance results, as demonstrated by five successful studies. These studies have created systems that are accurate and effective in meeting users' needs, with high precision in the results. The research also indicates that the VSM method is one of the best methods to apply in information retrieval systems due to its accuracy in delivering results.

References

Amin, F. (2012, June 18). Sistem Temu Kembali Informasi dengan Metode Vector Space Model. Jurnal Sistem Informasi Bisnis, 2(2). https://doi.org/10.21456/vol2iss2pp078-083

Amrizal, V. (2018, November 28). Penerapan metode term frequency inverse document frequency (tf-idf) dan cosine similarity pada sistem temu kembali informasi untuk mengetahui syarah hadits berbasis web (studi kasus: hadits shahih bukhari-muslim). Jurnal teknik informatika, 11(2), 149–164. https://doi.org/10.15408/jti.v11i2.8623

Anna, Hendini, A. (2019, June 26). Implementasi Vector Space Model Pada Sistem Pencarian Mesin Karaoke. Uinjkt. https://www.academia.edu/39702623/IMPLEMENTASI_VECTOR_SPACE_MODEL_PADA_SISTEM_PENCARIAN_MESIN_KARAOKE

Bahri, S. (2020, August 2). Aplikasi Pencarian Bahan Pustaka Di Perpustakaan Menggunakan Metode Vector Space Model. JIMP - Jurnal Informatika Merdeka Pasuruan, 5(2). https://media.neliti.com/media/publications/465266-none8c72b24a.pdf

Harna Yossy, E. (2020, June 15). Metode-Metode Information Retrieval | BINUS Online. BINUS Online. Retrieved October 18, 2023, from https://onlinelearning.binus.ac.id/computer-science/post/metode-metode-information-retrieval

Irmawati, I. (2017, May 1). Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model. Jurnal Ilmiah FIFO, 9(1), 74. https://doi.org/10.22441/fifo.v9i1.1444

Iswika, O. D., Sa'diyah, L., & Asep, A. (2022, June 24). Pengaruh Pemahaman Sistem Temu Kembali Informasi Pemustaka Terhadap Pemanfaatan OPAC (Online Public Access Catalog) Di UPT Perpustakaan Universitas Dehasen Bengkulu. LIBRARIA: Jurnal Perpustakaan, 10(1), 31. https://doi.org/10.21043/libraria.v10i1.13910

Nazya, M. F. (2017). KONSEP Customer Relationship Management (CRM) Pada Sistem Temu Kembali Perpustakaan Digital Metode Weight Adjusted K-Nearest Neighbor (WAK-NN) DAN Minimum Spanning Tree (MST) "STUDY KASUS PERPUSTAKAAN UIN SUSKA RIAU". Sistem Seleksi Proposal Tugas Akhir, 2(3).

Prabowo, Y. D., Marselino, T. L., & Suryawiguna, M. (2019, April 26). PembentukanVector Space Model Bahasa Indonesia Menggunakan Metode Word to Vector. Jurnal Buana Informatika, 10(1), 29 https://doi.org/10.24002/jbi.v10i1.2053

Siregar, R. R. A., Sinaga, F. A., & Arianto, R. (2017). Aplikasi Penentuan Dosen Penguji Skripsi Menggunakan Metode TF-IDF dan Vector Space Model. Computatio: Journal of Computer Science and Information Systems, 1(2), 171-186.

Triana, A., Saptono, R., & Sulistyo, M. E. (2014). Pemanfaatan Metode Vector Space Model dan Metode Cosine Similarity pada Fitur Deteksi Hama dan Penyakit Tanaman Padi. ITSMART: Jurnal Teknologi dan Informasi, 3(2), 90-95.

161–172.