Jurnal Elektronik Ilmu Komputer Udayana
Volume 9, No 1. August 2020

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Search Of File Journal With Query Word on List of Journal Document List Using Vector Space Model Method

Maula Khatami[1]
[1]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Bukit Jimbaran-Bali, Indonesia
[1]Khatamimaula@gmail.com

Abstract

Journals are articles about research that are very useful among academics and students alike. Every time we learn a new knowledge, we certainly need a guide that is verified and also credible. Students and academics were greatly helped by this journal. With journals help students and academics get references from previous research and get more insights so that they are able to make a related research and can even be improved from previous research. However, there are still many students and academics who find it difficult to find the right journal for their needs. So here the authors make a research system of information retrieval about journal searches by querying words using the vector space model method. In the suffix tree clustering method and the Vector Space Model, each document and keyword that has been carried out by the Text Mining process is then given the weight of each word contained in each existing document with the Term Frequency - Inverse Document Frequency (TF-IDF) weighting algorithm. ***Keyword:*** *Journal, Searching, Vector Space Model, Suffix Tree Clustering, TF-IDF*.

1. Introduction

Increased information flow that is very fast in supporting browsing and searching activities for users to facilitate their activities in getting information quickly, relevant, and according to the desired needs. This was followed by the development of Information Retrieval (IR) technology, which is a material search system (text documents) of unstructured properties (text) so that it is able to meet the information needs of a large set of documents (on a local computer server or the internet). In principle, information retrieval and information retrieval system is a simple matter, for example there is a place for storing documents (corpus) and the user formulates a question (request or keyword) whose answer is a collection of documents containing necessary information which expressed through user questions.

An alternative classification of documents based on the value of the level of similarity between existing documents and keywords entered and increase the level of relevance of the document's retrieval results into an information retrieval system, namely using the Suffix tree clustering algorithm and Vector Space Model, compared to other methods of classifying documents, the suffix tree clustering method and the Vector Space Model have several advantages, the value of ranking clearly in information retrieval, partial matching of keywords and also producing reference results that suit your needs.

In the suffix tree clustering method and the Vector Space Model, each document and keyword 1

that has been carried out by the Text Mining process are then given the weight of each word contained in each document that exists with the Term Frequency - Inverse Document Frequency (TF-IDF) weighting algorithm. The results obtained from the weighting of words of each document carried out the calculation of the measurement of the level of similarity by comparing the two corresponding vectors and then measuring the degree of similarity to keywords using the Cosine Similarity formula. Then we get a document classification results with the level of similarity that is close to the keywords.

2. Reseach Methods

The research method includes two methods namely, TF-IDF which is a comparison method and Vector Space Model which is a method for measuring the similarity between a document and a query.

2.1. Information Retrieval System

The Information Retrieval System is how to find a document from unstructured documents that provides the information needed from a very large collection of documents stored on a computer. The IR system accepts queries from users, then ranks documents in the collection based on their compatibility with the query. Ranking results given to users are documents that according to the system are relevant to the query. However, the relevance of documents to a query is a subjective user assessment and is influenced by many factors such as topics, timing, sources of information and user objectives. Here is the equation for calculating the number of clusters.

2.2. TF-IDF

TF-IDF is a term weighting method that is widely used as a comparison method for new weighting methods. In this method, the calculation of term weights in a document is done by multiplying the Term Frequency value with the Inverse Document Frequency. Term Frequency (TF) is a factor that determines the term weight in a document based on the number of occurrences in the document. Where every sentence, words will be sought to find the term, and the weight is calculated.

2.3. Vector Space Model

Vector space model is a model used to measure the similarity between a document and a query. Queries and documents are considered vectors in n-dimensional space, where t is the sum of all terms in the lexicon. The lexicon is a list of all terms that are in the index. Next will be calculated the cosine value of the angles of two vectors, namely W of each document and W of the keyword.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

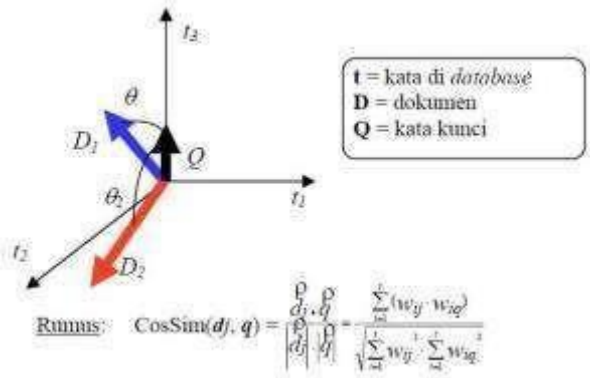

Figure 1. Vector Space Model

Vector space model is the solution to the problems faced when using the TF / IDF algorithm. Because in the TF / IDF algorithm there is a possibility between documents having the same weight, so it is ambiguous to be sorted. The Flowchart of the search uses the Vector space model algorithm as follows:

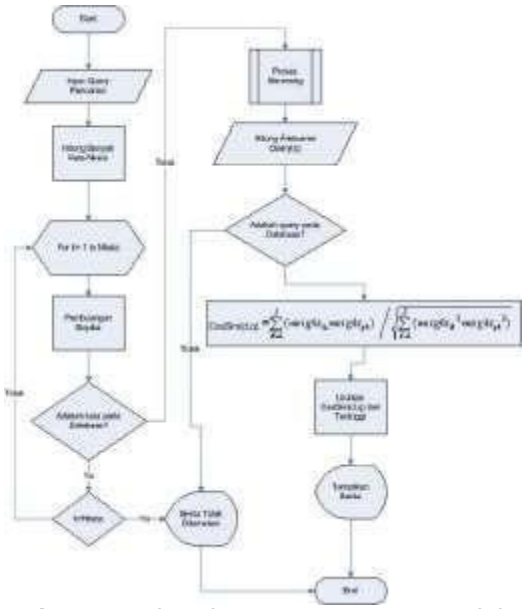

Figure 2. Flowchart Vector Space Model

3. Result and Discussion

The process begins with a database search, then goes into the preprocessing process, then enters the process of indexing words, after that the process of text mining, and finally measuring the similarity of the word.

3.1. Data Set

The data used in this study is in the form of research journal data in Indonesia. All Indonesian-language data and also query data will be testing data to calculate the accuracy of the system. In the final project that we made is still simple in the sense that the scope is still small, we will load a few .txt files from a computer that has been filled with several sample journals. From the program that we are going to run we will do a keyword search to find the document in which group of journals, for example the Journal of Informatics whether or not the Economic Journal and so on.

3.2. Document and Query Pre-Processsing

The stages start from processing unstructured data to structured to filtering data to find a knowledge or the relevance of the results of information needed by the user by the system, described as follows:

1) Users are required to enter online document links (at least 10 links), use of languages (Indonesian and English) that they want to use and keywords.
2) The system stores online document links and keywords that have been entered by the user into the database.
3) Text Mining Process, at this stage the Text Mining technique is performed on an online document that has been obtained from the previous step which will then be cleaned and prepared for the next stage. The process for preparing documents includes the process of cleaning up documents from HTML tags and unneeded characters, the process of deleting stopwords (conjunctions) and the process of stemming. The Text Mining process includes the Tokenization process, Stopword removal and Stemming process.
4) Interpretation and Evaluation Process, in this process patterns that have been identified by the system are then translated / interpreted in the form of knowledge that is more easily understood by the user to help in knowing the results given by the system or other forms that are easier to understand. Calculation of similarity in the previous step will produce a weight in each document that determines how relevant the document is to the query, so that only relevant documents can be displayed, in order starting from the most relevant (highest weight).
5) Process Testing (testing), in this process testing the results of each process carried out by the system. To obtain software with good results, it is necessary to measure the quality of the results obtained from the system:
   a) Testing of the results of the Text Mining process of the system.
   b) Tests on the results of the process of calculating the word weighting (Tf-Idf) and the results of measuring the level of similarity (measure similarity) of the system.
   c) Testing of computational time obtained by using a stopwatch and the difference in document size scale before and after the Text Mining process is performed by the system.
   d) Testing the evaluation results generated by the system by calculating the Recall value and Precision value to determine the optimal level of system results. Evaluation of an information retrieval system using Recall and Precision is good enough to be a measure of the system.
   e) Recall and Precision test testing is to obtain information on search results obtained by the information retrieval system created. Precision can be considered a measure of accuracy, while Recall is perfection. The value of Precision is the proportion of documents taken by the system is relevant. Recall value is the proportion of relevant documents

p-ISSN: 2301-5373
e-ISSN: 2654-5101

taken by the system. Recall and Precision values are between 0 and 100%. The information retrieval system is expected to provide Recall and Precision values close to 100% as the accuracy and perfection of the system in producing relevant information.

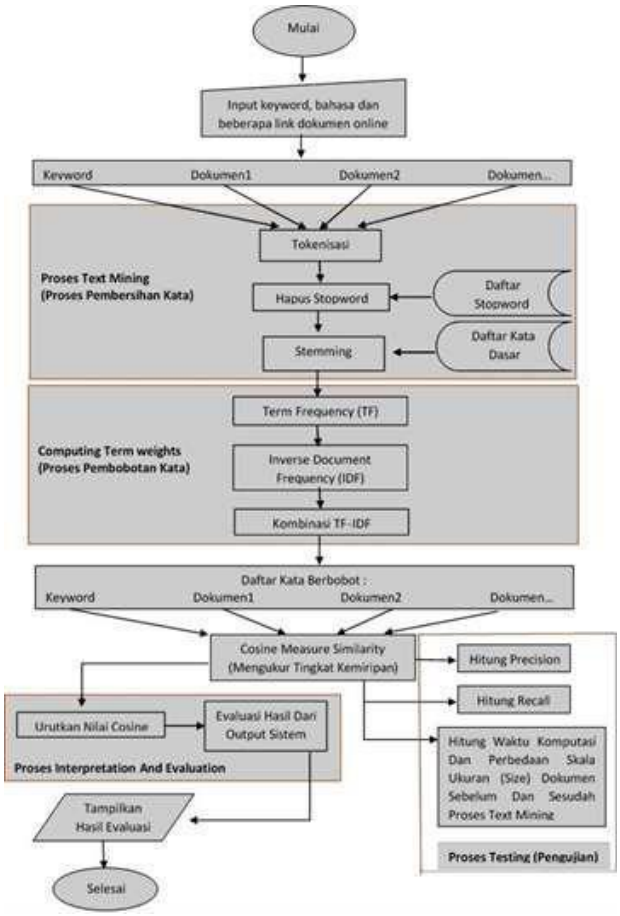

Figure 3. Pre-Processing

3.3. Index Modules

The indexing module is used only when the document retrieval system indexes the document. In this module parsing, stopwords and stemming process from documents to be indexed will be continued by calculating the value of variable frequency - invers document frequency (TF-IDF) and measuring similarity level. The results of indexing will be stored in a database that will be used when searching documents. The beginning of this indexing module is to save the link address and title of each online document that the user has entered into the database. Furthermore, the information will be retrieved documents that will be indexed from the database. Information taken is only based on the online document link address, not from the contents of the online document itself. The process of retrieving document information is based on the online document link address in the database, beginning with cleansing words in online documents from

unneeded characters, such as punctuation, symbols, html tags, javascript and others. The process of cleansing this word is done on every existing online document.

3.4. Text Mining Process Submodule

After the information about the document has been obtained based on the results of the previous word cleansing process, it will check the language used to retrieve information in the database that was previously entered by the user. The language checking process is important to separate the process of Text Mining with Indonesian and the Text Mining process with English because each language has different stopwords and different word stemming processes. The results of this submodule, namely indexing of the results of each stage in the Text Mining process. The interface of the indexing page in the Text Mining process.

After the information has been saved into the database in the form of a text file, then the information can be done calculation of weighting of words with Tf-Idf. Word weighting starts with calculating the weight of each word against the keyword and all information in the index database. The indexing is sorted starting from the keywords and then information from each stored database. The results of this submodule, namely indexing the steps and results in the process of weighting the word using the TF-IDF algorithm. The interface of the indexing page in the word weighting calculation process with the TF-IDF algorithm.

3.5. Measure Similarity Level Submodule

The final step in the indexing module when the word weighting calculation process has been completed, namely by calculating or measuring the level of similarity vector (information content) keywords with each information from each database. With the initial step, calculate the results of the scalar multiplication between the keyword weight value and each information weight from each other database. The result is the multiplication of each information weight with the keyword weight added. In the steps above, each calculation result obtained, both the scalar multiplication calculation and the length calculation of all information weights will be stored in the database.

Next is calculating the similarity between vector (information), from the keyword vector value with the vector value of each information in the order in the stored database. Calling of the results of the previous calculation is stored in the database needed to calculate the similarity between vectors. The results of this submodule, namely indexing of the results of measuring the degree of similarity of information with the formula Cosine Similarity in each document in the database on keyword information. The interface of the indexing page to the results of the measurement of the degree of similarity with the Cosine Similarity formula.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

```
In [1]: runfile('D:/VSM-master/vsm.py', wdir='D:/VSM-
master')

Masukkan Kata Yang Ingin Dicari >>> algoritma
Skor Kata :
0.107780254796
Kata tersebut ditemukan pada file : documents/
Jurnal_Matematika.txt


Skor Kata :
0.0789604682421
Kata tersebut ditemukan pada file : documents/
Jurnal_Teknik_Informatika.txt
```

Figure 4. Word Searching

Figure 5. Word Searching in File .txt

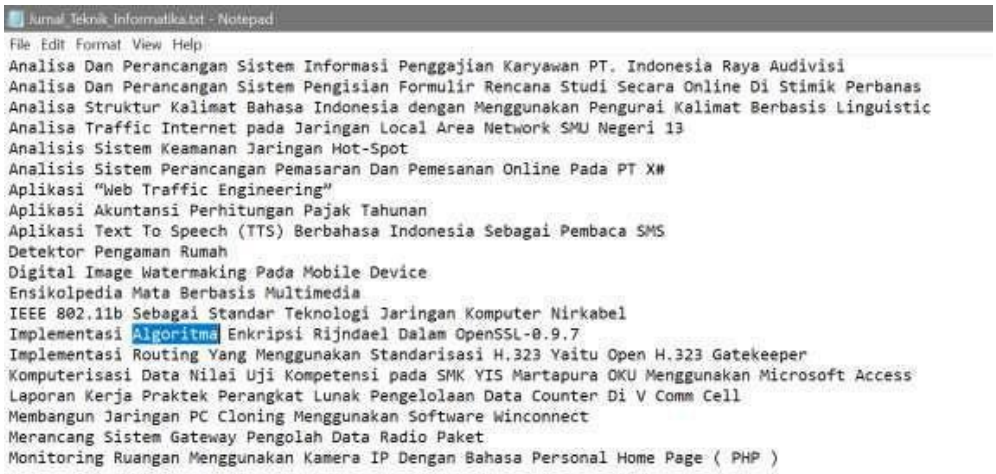

Figure 6. Word Searching in File .txt

In the search for 'algorithm' queries, the program shows that the journal documents sought are contained in two journals, namely in the Informatics Engineering Journal and the Mathematical Journal, and when checking, both documents contain the word 'algorithm', so the calculation of its accuracy:

- Precision (Relevant Data/ Retreive Total Data)
  Precision = 2 / 2 + 0 * 100
  = 100%

- Recall (Retrieve Data/ The Same Total Class Data in The Database)
  Recall = 2/2 * 100
  = 100%

4. Conclusion

From the results of the implementation of the program, the program can display the query / word search results in existing journal documents in the .txt document by the user, the closer to the value of 1 for the cosine score, the documents the documents sought by the user are more similar and there may be words we input there in several document files. On the implementation results only display documents with similarity above 0 and do not display documents with similarity below 0 or 0.

It is hoped that in the future this simple information retrieval system program can still be combined with other algorithmic methods and also developed to further improve the accuracy of the document search results desired by the user towards more complex.

References

[1] Sugara Bayu, Dody and Donny, "Sistem Temu Kembali Informasi Pada Gejala Autisme Dengan Metode Vector Space Model" *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) Vol. 3 No. 2 (2019) 257 – 264*.

[2] Fauzi Ahmad and Ginabila, "Information Retrieval System Pada File Pencarian Dokumen Tesis Berbasis Text Menggunakan Metode Vector Space Model" *Jurnal PILAR Nusa Mandiri Vol. 14, No. 2 September 2018*.

[3] Ridwan and S. A. Alfian Tomi, "Penerapan Mesin Pencari Informasi Dengan Menggunakan Metode Vector Space Model" *JUTEKIN Vol 7 No. 2 (2019) – PISSN : 2338-1477 – EISSN : 2541-6375*.

[4] Fauziah Siti, D. Nur Sulistyowati, Asra Taufik, "Optimasi Algoritma Vector Space Model Dengan Algoritma K-Nearest Neighbour Pada Pencarian Judul Artikel Jurnal" Jurnal PILAR Nusa Mandiri Vol. 15, No. 1 Maret 2019.