Information Retrieval System to Find Articles and Clauses in UUD 1945 Using Vector Space Model Method

O R Sulaeman¹, W Gata², E Wahyudi¹, M J Hakim¹, R Subandi¹, R Setiyawan¹, and B Pratama¹

¹Master of Computer Science Postgraduate Program, STMIK Nusa Mandiri, Indonesia
²Graduate School Master’s Degree Computer Science, STMIK Nusa Mandiri, Indonesia

Abstract. This study aims to find articles and clauses from the 1945 Constitution (UUD 1945) using the Vector Space Model method that calculates the similarity of many documents. One document is represented by one clause from each article of the 1945 Constitution. The next step is pre-processing by deleting unnecessary words (stopwords) and changing it into basic words (stemmer) in the Indonesian language. Each document will be indexed to speed up query and simplify the weighting. Words weighting in documents is performed using the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm by calculating the frequency of words in documents and all documents. The document search results will be presented in the ranking with the largest number of scoring appears at the top (descend sorting). The word search in this system more or less takes 90-100 milliseconds in 73 documents.

Keywords: UUD 1945, Constitution, Vector Space Model, Information Retrieval System.

1. Introduction

The 1945 Constitution of the Republic of Indonesia, or abbreviated as the 1945 Constitution, is a written basic law, the constitution of the current government of the Republic of Indonesia. The 1945 Constitution was ratified as a state constitution on August 18, 1945. During the period of 1999-2002, the 1945 Constitution had been amended 4 times. The amendment changed the composition of institutions within the constitutional system of the Republic of Indonesia.

Information retrieval is a method for rediscovering unstructured data stored in a set of documents and subsequently providing information about the subject needed. The purpose of this information retrieval system is to provide the information required by the users by getting all relevant documents that users need and also discard most of the irrelevant documents.

Users can find relevant information by reading all the documents in their storage, storing relevant documents, disposing of irrelevant documents, and sorting documents as needed. It is a perfect information retrieval system, but this solution is not practical and efficient because users do not have enough time to read all existing documents [1].

There are several methods of document information retrieval systems. They are the Gravitation Based Model, the Latent Semantic Model, the Vector Space Model, and the Generalized Vector Space Model. A good model of information search allows users to determine quickly and accurately whether the contents of the documents received are exactly what they need.

The Vector Space Model (VSM) is a technique used to represent documents and queries as vectors in multidimensional space. The dimensions are the terms used to build an index to represent the documents. It is the most widely used technique for information retrieval due to its simplicity; efficiency over extensive document collections, and it is very appealing to use. The effectiveness of the VSM depends
mostly on the term weighting applied to the terms of the document vectors. The VSM has three phases:
(1) the document indexing phase, where content-bearing terms are extracted from the document text;
(2) the weighting of the indexed terms to enhance retrieval of document relevant to the user; and
(3) ranking the document according to the query based on the similarity measure [2].
The implementation of the VSM has been used in various scientific fields such as Computational
Linguistics [3], Expert Systems [4], Medical [5], Knowledge-Based Systems [6], Data and Knowledge
Engineering [7], and so on.
Research on Information Retrieval System using VSM method has been carried out to search the text-
based thesis documents by Ahmad Fauzi and Ginabila [8]. Another research on the “final assignment”
search and document similarity calculations in abstracts have been conducted by Putri E, Martono D,
and Lis D [9].
This study aims to find articles and verses from the 1945 Constitution (UUD 1945) using the VSM
method. It performs the calculation on the similarity of a large number of documents. One document is
represented by a clause from each article of the 1945 Constitution. VSM performs a text/word search
statistically by calculating the weight value of the word for each document queried with TF-IDF formula
then calculate the similarity of values in the query and all documents stored with the Cosine Similarity
formula [10]. The research finds that the system is designed and built to produce relevant results,
efficient and easy to implement, and compatible with all operating systems using Java programming.

2. Methods
2.1. Data set
Data is obtained from Legal information network documentation of House of Representative of the
Republic of Indonesia website. The study utilizes 73 text files with the total file size of 32.6 KB.

2.2. Literature study
The literature study was carried out by studying the theories related to the object of the study that are
found in articles, journals, and books. These resources are used as the basis for this research to find the
theoretical basis of what has been previously carried out. The theories studied are Information Retrieval
System, Text Mining, VSM, Term Frequency-Inverse Document Frequency (TF-IDF), ECS Stemming,
Java Programming, and others subjects.

2.3. Data collection
The study was carried out using data from the 1945 Constitution. The data is then converted into articles.
One article is saved in a .txt format file.
Samples from each file are as follows:

| No. | Title | Words |
|-----|-------|-------|
| 1   | Pasal 1 | negara, indonesia, satu, bentuk, republik, daulat, ada, tangan, rakyat, laksana, undang, dasar, hukum |
| 2   | Pasal 2 | majelis, musyawarah, rakyat, anggota, dewan, wakil, daerah, pilihan, umum, atur, lebih, lanjut, undang, sidang, dikit, kali, lima, tahun, ibu, kota, negara, putus, suara, banyak |
| 3   | Pasal 3 | majelis, musyawarah, rakyat, wenang, ubah, tetap, undang, dasar, lantik, presiden, wakil, berhenti, masa, jabat |
| 4   | Pasal 4 | presiden, republik, indonesia, pegang, kuasa, perintah, turut, undang, dasar, laku, wajib, bantu, satu, orang, wakil |
| 5   | Pasal 5 | presiden, hak, rancang, undang, dewan, wakil, rakyat, tetap, atur, perintah, jalan |
| 6   | Pasal 6 | calon, presiden, wakil, orang, warga, negara, Indonesia, sejak, lahir, pernah, terima, lain, hendak, diri, khianat, mampu, rohani, jasmani, laksana, tugas, wajib, bagi, syarat, jadi, atur, lebih, lanjut, undang |
2.4. Data processing
Data processing in this study is illustrated in the flowchart diagram below:

![Data processing flowchart](image)

The explanation from Figure 1 is as follows:
- Reading, the process of reading the text file.
- Tokenizing, a process that aims to sort out strings from titles or keywords based on each of the composition words.
- Filtering, a process that aims to eliminate non-essential words, such as conjunctions, adverbs, and others.
- Stemming, a process that aims to change words into basic words by eliminating the first and final affixes.
- Indexing, the word indexing process from a collection of words in order to speed up data search.
- Term weighting is the process of weighing a word or term based on the number of frequencies and the existence of the word in the documents of system. The method used in the term weighting is Term Frequency-Inverse Document Frequency (TF-IDF) [11].

2.5. Data storing
The results of data processing are stored in memory that arranged by Map objects in Java. Data stored in memory enables faster searching compared to the data stored on the hard disk.

2.6. Information Retrieval System with a Vector Space Model
The indexing results are calculated by the level of similarity with the query using the VSM method. The steps are as follows:
- Stemming query, this process is similar to stemming in the data processing. This process aims to convert the query word into a basic word so that it can be found in indexing data.
- Calculating word weights in documents using TF-IDF (Term Frequency-Inverse Document Frequency) formula. IDF calculation is added to 1 so that the log value is not 0, so if multiplied by TF (total frequency), then the weight value becomes 0.
  \[ W_f = \frac{W_f}{T_f} \times \left( \log \left( \frac{D}{df} \right) + 1 \right) \]
  
  - $W_f =$ Term / word weight in the document  
  - $T_f =$ Number of query words in the document  
  - $D =$ Document total  
  - $df =$ Number of the document that has query word
- Calculating the distance from each document.
  \[ |Q| = \sqrt{\sum_{j=1}^{n} Q_j^2} \]
\[ |D| = \sqrt{\sum_{j=1}^{n} D_j^2} \]

Q = Query word  
D = Document word  

- Calculating dot product from query and document.
  \[ Q \times D = \sum_{j=1}^{n} Q_j \times D_j \]

- Calculating between query and the document using Cosine Similarity formula.
  \[ \text{Cosine } \theta = \frac{Q \cdot D}{|Q||D|} \]

- Ranking of documents uses the descending method (from the largest to the smallest similarity value). The document with the largest value means the document closest to the query.

3. Result and analysis

3.1. Data Indexing

During the processing of the files, each file is read and filtered by removing the stopword not required in the indexing. Furthermore, by stemming initial word becomes shorter.

Table 2. Word summary table (Top 40).

| No | Word   | Frequency | Doc Total |
|----|--------|-----------|-----------|
| 1  | undang | 177       | 49        |
| 2  | wakil  | 116       | 32        |
| 3  | presiden | 115   | 27        |
| 4  | rakyat | 87        | 33        |
| 5  | negara | 82        | 39        |
| 6  | dewan  | 70        | 26        |
| 7  | hak    | 68        | 20        |
| 8  | dalam  | 62        | 34        |
| 9  | oleh   | 55        | 26        |
| 10 | 2      | 53        | 49        |
| 11 | daerah | 50        | 12        |
| 12 | 1      | 50        | 49        |
| 13 | atur   | 49        | 35        |
| 14 | tiap   | 43        | 24        |
| 15 | dapat  | 41        | 23        |
| 16 | pilih  | 40        | 12        |
| 17 | 3      | 40        | 34        |
| 18 | serta  | 37        | 24        |
| 19 | orang  | 35        | 16        |
| 20 | anggota| 34        | 15        |

Table 2 is a Top 40 summary of indexing on stored words. In Word column are columns of stemming words. Frequency column is a column for the number for the frequency of certain words in all documents. Document Total column is a column for the number of documents that have a particular word. The number of words stored in indexing from 73 documents is 486 words.

3.2. Results of Query

Query testing is three steps by increasing the number of query words for each step as follows:

- Documents that match the query.
- A score of the similarity of the query with the document.
- The time needed to search.

3.2.1. The First Testing

This test search for a query word “mahkamah” in the search system. The test results are as follows:
3.2.2. The Second Testing
This test search for two query words “mahkamah” and “konstitusi” in the search system. The test results are as follows:

![Figure 3. Second testing result.](image)

3.2.3. The Third Testing
This test search for three query words “mahkamah”, “konstitusi”, and “hakim” in the search system. The test results are as follows:

![Figure 4. Third testing result.](image)

From the three tests above, the comparison of query time and documents total are as follows:

| Testing | Query Time (ms) | Document Total |
|---------|----------------|----------------|
| 1       | 93             | 6              |
| 2       | 94             | 6              |
| 3       | 99             | 8              |

Table 3. Comparisons testing
3.3. Results of Query
Based on the three tests above, the more query words are displayed, the more proper documents in the retrieval system (3rd test) appear. It happens because the search system will look for documents that have a minimum relationship of one query word. More query words mean more documents related to the query.

The number of query words does not influence the score because the score depends on the suitability of the query word with the document. The score will be higher when the document relates with more query words. The lowest score means that the document only has 1 query word. The number of query words has no effect on the search time in the system because the indexing data has been stored in memory. If the search system adds more documents, the longer the search time will take, and the slower the process because the number of objects in memory is higher.

4. Conclusion
From the research have done can be concluded as follows:

1) The application of the Information Retrieval System uses VSM method can be done for Article and Clause in the 1945 Constitution.
2) The retrieval system can search for information from the contents text files quickly without being affected by a large number of query words.
3) The stemming process of keywords makes it easy to search between words in the query and in the document.
4) The VSM method calculates similarity based on the statistical value of the words in the document, so it has the disadvantage of searching interrelated words or phrases.

There are many methods for the retrieval system; it is expected that the researcher can use the method for the same object and compare the best retrieval method in that case.

5. References
[1] Jasman Pardede, Mira Musrini, and Wildan Denny P 2013 *Implementasi Metode Generalized Vector Space Model pada Aplikasi Information Retrieval*. Jurnal Informatika, No.1 Vol. 4, pp. 57-68.
[2] E E Ogheneovo and R B, Japheth 2016 *Application of Vector Space Model to Query Ranking and Information Retrieval* International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 5, pp. 42-47.
[3] K. Erk, S. Padó, and U. Padó 2010 *A flexible, corpus-driven model of regular and inverse selectional preferences*. Computational Linguistics, Volume 36, Issue 4.
[4] K Y Bum and K Y Seopc 2010 *An autonomous assessment system based on combined latent semantic kernels* Expert Systems with Applications: An International Journal, Volume 37 Issue 4, pp. 3219-3228.
[5] C Lopez and C Ribeiro 2010 *Using Local Precision to Compare Search Engines in Consumer Health Information Retrieval* SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval.
[6] B Yu and D Zu 2009 *Combining neural networks and semantic feature space for email classification Knowledge-Based Systems*, Volume 22, Issue 5, pp. 376-381. July 2009.
[7] W Mao and W W Chu 2007 *The phrase-based vector space model for automatic retrieval of free-text medical documents* Data & Knowledge Engineering Volume 61, Issue 1, pp. 76-92, April 2007.
[8] Ahmad Fauzi and Ginabila 2018 *Information Retrieval Sistem pada File Pencarian Dokumen Tesis Berbasis Text Mining Menggunakan Vector Space Model* Jurnal PILAR Nusa Mandiri Vol. 14 No. 2 pp. 41-46.
[9] Putri E, Martono D, and Lis D 2017 *Information Retrieval Tugas Akhir dan Perhitungan Kemiripan Dokumen Mengacu pada Abstrak Menggunakan Vector Space Model*. Jurnal SIMETRIS, Vol 8 No 1, April 2017.
[10] D L Lee, H. Chuang, and K Seamons 1997 *Document Ranking and The Vector-Space Model* IEEE Software Volume: 14, Issue: 2, DOI: 10.1109/52.582976, Mar/Apr 1997.
[11] P Soucy and G W Mineau 2005 *Beyond TFIDF Weighting for Text Categorization in the Vector Space Model* IJCAI'05 Proceedings of the 19th international.