Performance comparison of CATA’s information retrieval system application vs WTS

Jajang Jaya Purnama¹, Sri Rahayu², Sfenrianto³, Fitra Septia Nugraha⁴, Ridan Nurfalah⁵, Hendri Mahmud Nawawi⁶, Muhammad Ja’far Shidiq⁷, Nurul Ichsan⁸, Haerul Fatah⁹

¹²⁴⁵⁶⁷ Master of Computer Science-Postgraduate Program, STMIK Nusa Mandiri, Jakarta, Indonesia
³Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management, Bina Nusantara University, Indonesia
⁸⁹ Information Systems, Bina Sarana Informatika University, Indonesia

E-mail: ¹jajangja2412@bsi.ac.id, ²srahayu110527@gmail.com,³sfenrianto@binus.edu, ⁴fitraseptia7@gmail.com, ⁵ridannurfalah1995@gmail.com, ⁶mahmudhend94@gmail.com, ⁷ash.shidiq.mj@gmail.com, ⁸nurul.nrc@bsi.ac.id, ⁹haerul.hef@bsi.ac.id

Abstract. Searching for data or documents that are always a necessity in the world of education, work, research others requires convenience so that time efficiency occurs. This is supported by the emergence of various Information Retrieval system applications, one of which is cata, which is built on the java programming language and applies the Generalized Vector Space Model method which can search based on similar keywords. Other facilities need to be developed, one of which is by applying the stemmer porter algorithm applied to the WTS (Web Tokenization Stemming) web which is built with the laravel framework that can search data or documents simply by typing the basic words so that searches are easier and faster.

1. Introduction
Applications for information retrieval or document are important in the development and application of information retrieval systems (IRS). This system is useful for finding information that is usually in the form of documents from unstructured data in the form of text [1]. Currently developers are developing Information Retrieval Systems with various methods, such as fuzzy logic [2], vector space models (VSM) [3], neural networks [4], and various other methods. One example of an IR System that applies vector methods, namely the Generalized Vector Space Model (GVSM) [5], which is then implemented in the CATA application [6].

This application serves to process queries and serves to make matching between queries with words that exist in the document [6]. Then, sort the documents that are relevant to the query based on the similarities between the highest queries and documents [7]. Thus, the search algorithm is based on keywords that must be similar.

The ease of searching is needed in information retrieval [8]. Porter Steemer algorithm is a computational procedure that converts words into their original form (stem) by searching for prefixes, suffixes and deleting them according to the rules of a language [9]. This algorithm is applied by
compilers in a web, hereinafter referred to as Web Tokenization Stemming (WTS), which can search based on basic words, making it easier for users to conduct searches [9]. Thus this study aims to propose an IRS application using the CATA and WTS approaches.

2. Literature Study
Various discussions that support this research are very broad but are limited by various literature that is used and supports this research, including the following:

2.1. Information Retrieval System
Information Retrieval (IR) system is used to retrieve information that is relevant to the needs of users of a collection of information automatically.

One common application of the IR system is the search engine or search engine found on the internet network. Users can search the web pages that they need through the engine direction. Another example of the IR system in the library information system[10].

2.2. Generalized Vector Space Model
Generalized Vector Space Model (GVSM) is an extension of the Vector Space Model (VSM) by adding additional types of information, in addition to terms, in representing documents, IRS with GVSM represent document with vector similarity to all existing documents[11].

2.3. Cata Application
This application is an example of an IR system that applies vector methods, namely the Generalized Vector Space Model, which is then called the Cata application. This application serves to make matching between queries with words that are in the document. This application displays documents that are relevant to the query and sort them based on the similarities between the highest queries and documents. In this application, there is no access category for users, so all users can use this application. This application is needed to make it easier for users to find information in the document. There are also features to add, change and delete documents in the document collection[10].

2.4. Porter Steemer Algorithm
A stemmer is built on the assumption that there is no word that is double. The process carried out is to remove particles, eliminate the pronouns, eliminate the first prefix, eliminate the second prefix, eliminate the suffix and then find the basic words of a word[9].
2.5. WTS (Web Tokenization Stemming)
Web Tokenization Stemming or WTS is an Information Retrieval System website that was built using the Ralafel framework by applying the Porter Steemer Algorithm which aims to make it easier for users to search by simply giving the basic word.

3. Research

3.1. Research Object
The author conducted a study of several primary documents that the author has in the form of several primary documents that the author has in the form of PDF which IRS will process using the Porter Algorithm and compared with the performance of the GVSM algorithm in the Cata application.

3.2. Research Method
1. Determining Data
   The author uses the data for this study is the primary data that belongs to the author in the form of PDF, which will later be processed according to Porter’s algorithm.
2. Literature Study
   After the author obtains the data, then searches for literature to make a benchmark for the research conducted.
3. Stemming Method
   The method chosen from the existing problems, a stemmer is built on the assumption that there is no word that is double.
4. Tokenization
   The process of diving text can be sentences, paragraphs or documents into tokens.

4. Result and Discussion
Users can access the IRS CATA and WTS applications to find documents needed on the database. The search uses the data CATA application that uses the Generalized Vektor Model (GVSM) and Web Tokenization Stemming (WTS) that uses the Porter Steemer algorithm. Search results in the form of documents needed by the user based on basic words.

![Figure 2. The Architecture Application of IRS CATA-WTS](image)

4.1. GVSM in IRS Cata’s Application
For example, there is a query (Q), and 3 documents, namely document 1 (D1), document 2 (D2), and document 3 (D3) as follows:
Q: Penyelesaian konflik Aceh
D1: Gus Dur Tak Mungkin Dijatuhkan
D2: Bondan Bukan Saya Nggak Doyan Duit
D3: AS Dukung Kesepakatan GAM-RI
Form the results of the search in the journal Cata Information Retrieval (IR) Application with Generalized Vector Space Model the similarity model above is taken ranking is document 1, document 3, document 2 which means document 1 is most relevant to the query. This Cata application functions to process queries and serves to make matching between queries and words in the document. This application displays documents that are relevant to the query and sort them based on the similarities between the highest queries and documents[10].

Figure 3. Main Form of Cata Application

Figure 4. Structure of Stemming
4.2. Plot Tokenisation and Stemming Method (Porter Steemer Algorithm)

A stemmer is built with 3 foundations that are not ambiguous or meaning two, the stemming process is eliminating the prefix and suffix, eliminating the suffix and finding the basic word of a word. Figure 4 describes a process of steaming where stemming is removing the prefix and also the end of this structure begins by reading a word that is inputted, word input can be one or more, after reading remove particle, after particle the process of removing positive pronoun, after being fulfilled then remove the 1 order prefix, then the decision process, the successful decision will lead to the right, when the process fails to lead to the left, the two decision lines do not stop immediately, when the process successfully proceeds to the remove suffix, reunite with the decision if will successfully remove 2 order prefix and at the end of the process the steaming process will be carried out as the flow mark is fulfilled [9].

4.3. Website IR with Metode Tokenisation dan Steeming (Porter Steemer Algorithm)

This application is one of the retrieval applications that applies the tokenization system and stemming using the Porter algorithm, which is then the WTS website (Website Tokenisasi dan Stemming). This website serves basic word searches for submission, pdf file uploads, word searches, and tokenization tables. This website can be used by all users. The following is the index view of WTS (Website Tokenisasi Stemming)

![View of Index WTS](image1)

![Website of Nusa Mandiri](image2)
In Figure 5, in the image the user is given several navigation options and the main part is the group name of the author, the navigation is among others STMIK Nusa Mandiri, About, Basic Words, PDF File Upload, Word Search, Tokenisation Table. Figure 6 is the website of the affiliate of the author, STMIK Nusa Mandiri, which is addressed at Kramat Raya Road no.18, 01/07 Kwitang, Seneng Central Jakarta City, DKI Jakarta 10420.

![Figure 7. About WTS](image)

Figure 7. Displaying from the research title about information retrieval, campus name, and class code.

![Figure 8. Basic Word Search](image)

Figure 8. displaying basic word search, its use to search for basic words, for example, downloading basic words to download, we can enter a few words to find the basic word.

![Figure 9. View of Upload](image)
5. Conclusion
With the results of the comparative performance of the Cata Applications and WTS (Web Tokenization Stemming), both of them are applications for supporting Information Retrieval System but both use different algorithms and different platforms. The cata application uses a desktop platform, WTS uses a website platform. The GVSM algorithm is used for the Cata application with conclusions using the weight of the index term, the existence of document vectors and queries, cross-product calculations to determine the similarity of queries and documents. While the porter algorithm is used for WTS using the stemming and tokenization process which can use basic words so that it is easier to search.

References
[1] Adams, S. (2010). The text, the full text, and nothing but the text: Part 1–Standards for creating textual information in patent documents and general search implications. World Patent Information, 32(1), 22-29.
[2] Gupta, Y., Saini, A., & Saxena, A. K. (2015). A new fuzzy logic-based ranking function for efficient information retrieval system. Expert Systems with Applications, 42(3), 1223-1234.
[3] Liu, X., Gao, J., He, X., Deng, L., Duh, K., & Wang, Y. Y. (2015). Representation learning using multi-task deep neural networks for semantic classification and information retrieval.

[4] Abualigah, L. M. Q., & Hanandeh, E. S. (2015). Applying genetic algorithms to information retrieval using vector space model. *International Journal of Computer Science, Engineering and Applications, 5*(1), 19.

[5] Waitelonis, J., Exeler, C., & Sack, H. (2015). Linked data enabled generalized vector space model to improve document retrieval. *NLP & DBpedia@ ISWC, 15*, 34-44.

[6] Bunyamin, H., & Negara, C. P. (2008). Aplikasi Information Retrieval (IR) CATA Dengan Metode Generalized Vector Space Model. *Jurnal Informatika, 4*(1), 29-38.

[7] Bunyamin, H (2005). Information Retrieval System Dengan Metode Latent Semantic Indexing. *Institut Teknologi Bandung, Bandung*

[8] Lüke, T., Schaer, P., & Mayr, P. (2012, September). Improving retrieval results with discipline-specific query expansion. In *International Conference on Theory and Practice of Digital Libraries* (pp. 408-413). Springer, Berlin, Heidelberg.

[9] Bagus, P., Wiguna, S., & Hantono, B. S. (2013). *Peningkatan Algoritma Porter Stemmer Bahasa Indonesia berdasarkan Metode Morfologi dengan Mengaplikasikan 2 Tingkat Morfologi dan Aturan Kombinasi Awalan dan Akhiran.* 2(2), 1–6.

[10] Bunyamin, H. (2008). Information Retrieval System Dengan Metode Latent Semantic Indexing. *Jurnal Informatika, Vol 4, No.*, 29–38.

[11] Pardede, J. (2014). *gvsm.pdf. Jurnal Sistem Komputer, Vol 4 No 1,* 1–6.