Journal Basic Science and Technology 12(2) (2023) pp. 47-51

Published by: IOCSCIENCE

Journal Basic Science and Technology

Journal homepage: www.iocscience.org/ejournal/index.php/JBST

Implementation of Information Retrieval System for Document Search Based on User Keywords

Adjie Arya Sandjaya W[1], Sonia Ayu M.P[2], Nabila Calista S[3]
Department of Informatics Engineering, Bhayangkara University, Surabaya

ARTICLE INFO

Article history:

Received Aprl 19, 2023
Revised Mey 20, 2023
Accepted June 22, 2023

Keywords:

Document Indexing;
Document Relevance;
Information Retrieval System;
VSM.

ABSTRACT

This journal discusses the importance of information in everyday life and the challenges of finding specific information in the digital era where large amounts of data are available. An effective information retrieval system is needed to assist users in finding the information they need. This paper focuses on the implementation of an information retrieval system for searching news articles based on user keywords using the Vector Space Model (VSM) algorithm, which is commonly used in information retrieval systems. The system indexes and analyzes news articles using VSM to generate a list of relevant documents based on user keywords. The goal of this research is to explain how this system can help users find news articles that meet their needs and to discuss the advantages and disadvantages of information retrieval systems, as well as potential future developments in this field. This paper provides valuable insights for researchers and practitioners in the field of information retrieval systems.

Corresponding Author:

Adjie Arya Sandjaya Wardana,
Department of Informatics Engineering,
Bhayangkara University,
A. Yani 114 Wonocolo Road, Surabaya, 60231, Indonesia.
Email: aji17902@gmail.com

1. INTRODUCTION

Computer technology is currently developing rapidly. Affordable costs with high specifications are capable of addressing problems ranging from the simplest to the most difficult(Fadlil, 2018). Information has become one of the essential needs in everyday life. In the digital era like today, information can be easily and quickly accessed through the internet. However, with the abundance of available information, searching for specific information can be a challenge. Therefore, an effective information retrieval system is needed to assist users in finding the required information.

An information retrieval system is designed to search for documents or information that matches the criteria entered by the user. One form of information retrieval system is searching for news documents based on user keywords. This system allows users to search for relevant news on specific topics by entering appropriate keywords or phrases.

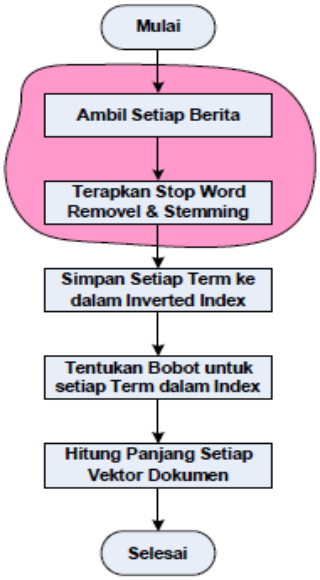
The difference between information retrieval and data search is that information retrieval primarily deals with searching for unstructured information, where the desired or relevant information is found based on keywords(Putung et al., 2016).

This journal will discuss the implementation of an information retrieval system for searching news documents based on user keywords. The method used is the Vector Space Model (VSM) algorithm, which is one of the most commonly used methods in information retrieval systems. In this

implementation, news documents will be indexed and analyzed using VSM to generate a list of documents relevant to the keywords entered by the user.

The objective of this journal is to explain how the implementation of an information retrieval system can help users search for news documents that meet their needs. Additionally, this journal will also discuss the advantages and disadvantages of information retrieval systems and how this system can be developed in the future. It is expected that this journal will provide valuable insights for researchers and practitioners in the field of information retrieval systems

2. RESEARCH METHOD



Ficture 1. Preprocessing, indexing, term weighting determination, and calculation of vector length for each document.

In general, the information retrieval system consists of two main stages, namely indexing and retrieval. After collecting the documents, the first step is indexing, which involves building an index list. The indexing process includes tokenization, removal of stop words, and stemming. Each important term is then stored in the index list.

The preprocessing stage also includes the removal of stop words, resulting in a more complex list of words or terms that still represent the processed documents. The list of stop words is stored in an array. Furthermore, the stemming stage is used to map and reduce words to their base form. This process transforms words with affixes into their base form. The list of base words for a term is stored in a table.

Weighting is also part of the indexing process. The weighting scheme used is tf.idf, where idf is calculated as log(n/N), where N is the number of documents containing a certain term, and n is the total number of documents in the collection (corpus).The indexing process involves four stages. First, the removal of stop words and stemming are performed. Then, each term resulting from the removal of stop words and stemming is stored in the index table. Next, the weights of each term in the index table are determined. Finally, the vector length of each document represented by the terms in the index table is calculated.

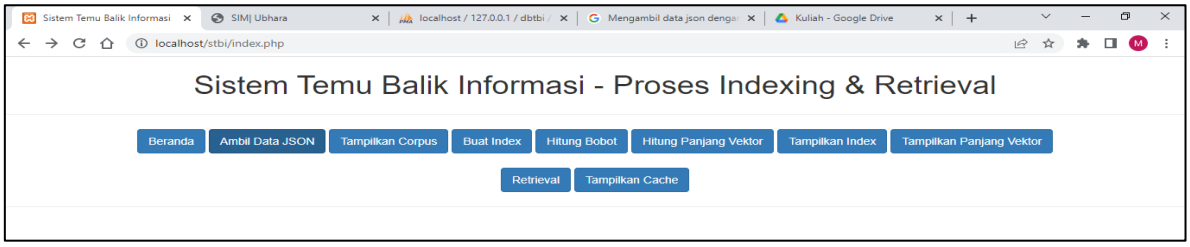
The retrieval stage begins with taking the user's query. The query is then subjected to the removal of stop words and stemming, resulting in more complex keywords that still represent the query. Based on these keywords, the system looks into the cache. If the keyword is already in the cache, the system will directly sort the documents similar to the keyword and return them to the user. However, if the keyword is not yet in the cache, the system will calculate the similarity between the keyword and the list of documents represented by the terms in the index. The calculation results are then stored in the cache table.

3. RESULTS AND DISCUSSIONS

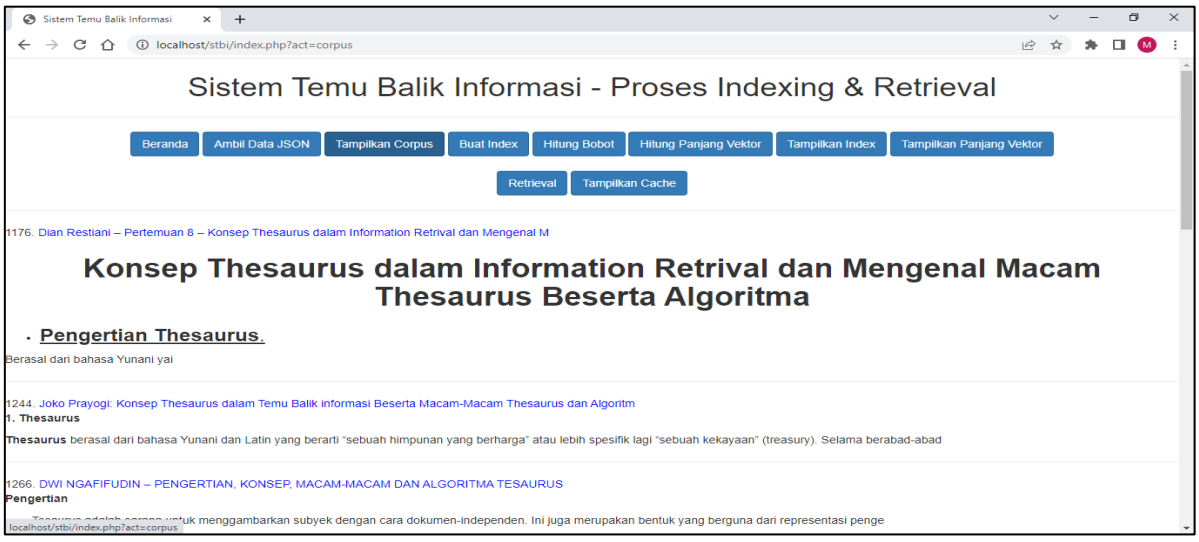
It starts by retrieving the documents from the database. Next, stop word removal and stemming are performed on the documents. After that, each term is stored in the inverted index. The weight for each term in the index is determined. Then, the length of each document vector is calculated. This process is illustrated in the figure below.

3.1. User Interface of the Retrieval System Application

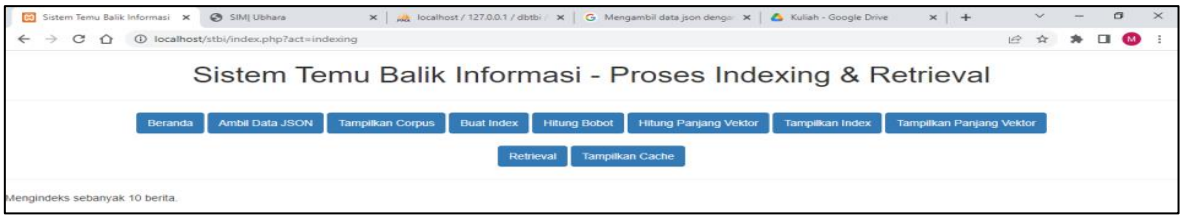
Ficture1. Home page of the information retrieval system application.



Ficture2**.** Retrieving JSON data from the database..



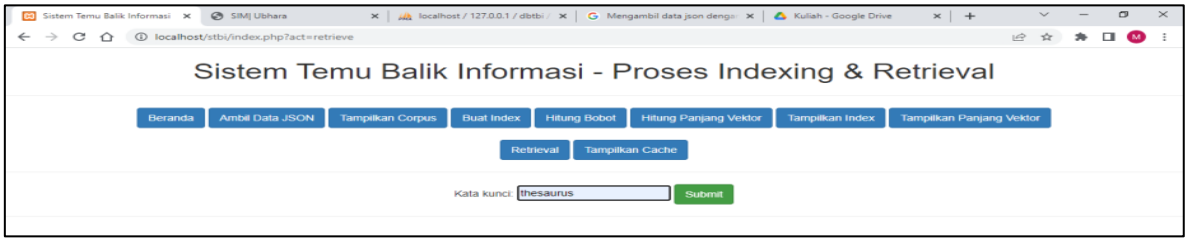
Ficture 4**.** Corpus view in the information retrieval system application.



Ficture 5. Creating an index in the information retrieval system application

Ficture 6 Calculating weights in the information retrieval system application

Ficture 7. Calculating vector length in the information retrieval system application



4. CONCLUSION

The implementation of an Information Retrieval System for Document Search based on User Keywords is a research study aimed at developing a system that can search and discover relevant news documents based on user-entered keywords. This method utilizes natural language processing techniques to identify relevant keywords within the document text. The study employs various techniques such as document indexing, natural language processing, and keyword search models to construct an efficient and accurate information retrieval system. This system assists users in quickly finding specific and relevant news documents without having to manually read through all the documents.With the implementation of this information retrieval system, the research concludes that users can easily search for the desired news documents based on the entered keywords. This system aids in managing and searching large-sized news documents, enabling users to obtain relevant information more efficiently.

REFERENCES

Afandi, S., Ardiansyah, F., & Soedarsono, B. (2016). Pengembangan Sistem Temu Kembali Informasi Digital Fulltext Artikel Jurnal Di Pdii – Lipi. *Baca: Jurnal Dokumentasi Dan Informasi*, *36*(1), 65. https://doi.org/10.14203/j.baca.v36i1.203

Christioko, B. V. (2012). Implementasi Sistem Temu Kembali Informasi Studi Kasus: Dokumen Teks Berbahasa Indonesia. *Jurnal Transformatika*, *10*(1), 1. https://doi.org/10.26623/transformatika.v10i1.64

Fadlil, A. (2018). Aplikasi Sistem Temu Kembali Angket Mahasiswa Menggunakan Application of Information Retrieval for Opinion Student. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *6*(1), 33–40. https://doi.org/10.25126/jtiik.201961184

Kadafi, A. R. (2018). Implementasi Sistem Temu Kembali Informasi Pada Dokumen Mutu. *Jurnal ELTIKOM*, *2*(1), 18–25. https://doi.org/10.31961/eltikom.v2i1.38

Prabowo, T. T. (2021). Efektivitas Sistem Temu Kembali Informasi Perpustakaan Digital Institut Seni Indonesia (ISI) Yogyakarta dalam Tinjauan Recall dan Precision. *Media Pustakawan*, *28*(1), 37–48. https://doi.org/10.37014/medpus.v28i1.1087

Putung, K. D., Lumenta, A. S. M., & Jacobus, A. (2016). Penerapan Sistem Temu Kembali Informasi Pada Kumpulan Dokumen Skripsi. *Jurnal Teknik Informatika*, *8*(1). https://doi.org/10.35793/jti.8.1.2016.12227