ATLANTIS PRESS

*Advances in Computer Science Research, volume 96*

**Proceedings of the International Conference on Mathematics, Geometry, Statistics, and Computation (IC-MaGeStiC 2021)**

# Information Retrieval Using Matrix Methods
## Case Study: Three Popular Online News Sites in Indonesia

Ferry Wiranto*, I Made Tirta

*Department of Mathematics, FMIPA, University of Jember*
**Corresponding author. Email: ferrywr25@gmail.com*

**ABSTRACT**
This research is part of data mining, a sub-section of information retrieval and text mining. This research focuses on finding an approach to getting relevant documents online news documents with a specific threshold value and improving computing performance to get relevant documents with large documents. In this case, the author use news from 3 news sites that are pretty popular in Indonesia, which are included in the top 10 Alexa Traffic Rank (ATR) 2021, namely tribunnews.com, detik.com, and liputan6.com. In searching for relevant news documents, the author determines the threshold value first by calculating the average similarity value of the documents used as the experimental sample. The resulting threshold value is a determinant of the similarity value of each document to be used. The author uses several techniques to assist the research process, such as text mining with the tala method and news document representation techniques using matrix methods, and finally utilizing the cosine size method to determine the similarity of documents with matrix-based search data. The results obtained indicate that the approach using the matrix method and the matrix compression process shows good computational results, so it will be useful for implementation on a large number of documents.

***Keywords:*** *Data mining, Text mining, Matrix methods, Cosine size, Sparse matrix*

## 1. INTRODUCTION

Website is one of the technologies currently developed to disseminate information very quickly and reachable. Secretary-General of APJII (Association of Indonesian Internet Service Providers) Henri Kasyfi Soemartono explained that the penetration of internet users in Indonesia rose from 64.8% in 2018 to 73.7% in 2020 of the population or the equivalent of 196.7 million users[1]. The increase in the number of users was partly due to several factors, such as fast internet or broadband infrastructure in Indonesia being more evenly distributed with the Palapa Ring, increasingly massive digital transformation due to online learning, and the policy of working from home due to the Covid-10 pandemic since last March. APJII conducted a survey method in 2020 using sampling techniques such as probability sampling, multistage random sampling, and area variant random sampling

Based on The results of the 2020 databox survey, The percentage of Indonesian people who access the most information through social media, television, online news or online news sites, and official government websites[2]. It was found that it can be concluded that in the current era, Indonesian people tend to prefer to seek information through their respective digital media, such as using mobile phones and laptops. Based on the 2020 databox survey results, we also find that the three media they access can be viewed through the website because online websites tend to be more flexible and easily accessible from various lines. In addition, the Importance of the website as a means of disseminating information is also shown by the number of visitor (traffic).

The development of the website as a form of mass media resulted in a sharp increase in the amount of information in news articles. The results of observations from three news sites (namely Tribunnews.com, Detik.com, and Liputan6.com) using the scraping technique, from January to October 2021, found as many as 210,997 news that has been published on three news sites used as author references. Authors use Tribunnews.com, Detik.com, and Liputan6.com because the news site in 10 most popular news sites in Indonesia in 2021[3], and the data is already available by previous research[4]. When viewed in terms of numbers, newsreaders are sufficient to get contacts from the information they want to know.

However, readers also have difficulties knowing the suitability of news documents with the news documents they are looking for with this amount of information. Based on previous research, The use of clickbait article titles is widely used by online media to attract readers' interest by disturbing their curiosity[5], so that news readers have difficulty finding relevant news. Readers must first read the news documents generated from the search engine and then manually check whether the news follows the data sought and the information follows information from other news sites. So Make the readers have difficulty and take a long time to find out and find the truth of information through news documents that match what they are looking for.

On the other hand, currently, the implementation of the retrieval system generally uses TF-IDF (*term frequency-inverse document frequency*). TF-IDF takes longer processing time than SQL queries so that SQL queries have a better efficiency level than the TF-IDF Algorithm[6]. So that the use of TF-IDF is still not optimal when used on big data. We propose an alternative technique to improve the computational performance for returning relevant news documents. The alternative technique uses a matrix approach along with a matrix compression process, so that later it is expected to speed up the computational process even though it is used on relatively large documents.

Information Retrieval is a science that studies procedures and methods for retrieving stored information from various relevant sources from a collection of information sources sought or needed. Several data types can be found in data search, including text, tables, images, videos, and sounds. Information Retrieval aims to fulfill user information by rediscovering relevant documents or reducing irrelevant search documents. In general, the flow of information retrieval can be seen in Figure 1.

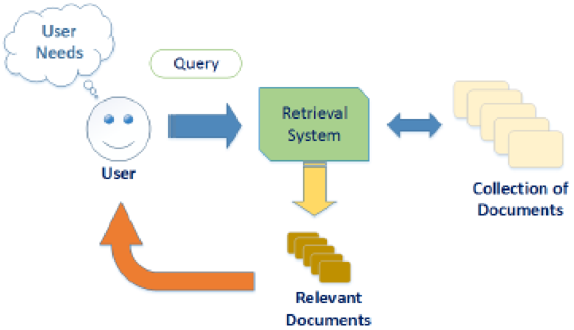


**Figure 1** Information retrieval flowchart [7].

Based on Figure 1, the process of displaying relevant documents to the user by first reading the query, followed by a system that uses an information retrieval model, then the query will be checked against the stored document, then returned. When a copy from the Collection of Document is obtained, the system will review it to see how far the paper is similar to the query. Then the system will display the relevant document according to the user query.

Text mining is a variation of the sub-field of data mining that seeks to find patterns or interesting data characters from a large set of textual data. In the process of steps in text mining, there is an initial processing stage of a text called text preprocessing. Text preprocessing is a process of converting unstructured textual data into structured data. This preprocessing process includes four steps, namely case-folding, tokenizing, filtering, and stemming: (i). Casefolding is the initial stage of the text mining process, which changes all the characters in the data into lowercase letters, (ii) Tokenizing Is the stage of cutting the input string based on each word that composes it. Tokenizing breaks down a set of characters into word units. The cuts connect one word and another in a sentence or a complete document, (iii) Filtering is taking essential words from the tokens generated. This process can use a stoplist algorithm to remove fewer necessary comments, (iv) Stemming is needed to reduce the number of different indexes of a document based on the wording of the compiler of the document. Stemming will eventually get basic words or words not included in conjunctions and adverbs. Lately, there are many text stemming techniques to help get essential words in a text. Previous studies also compared several stemming algorithms such as the Nazief and Andriani algorithm, the vega algorithm, the Arifin, Setiono algorithm, and tala algorithm[8]. However, The author used the tala algorithm in this study because the computation time is relatively faster, with more than 75% accuracy. An overview of process A can be seen in Figure 2.

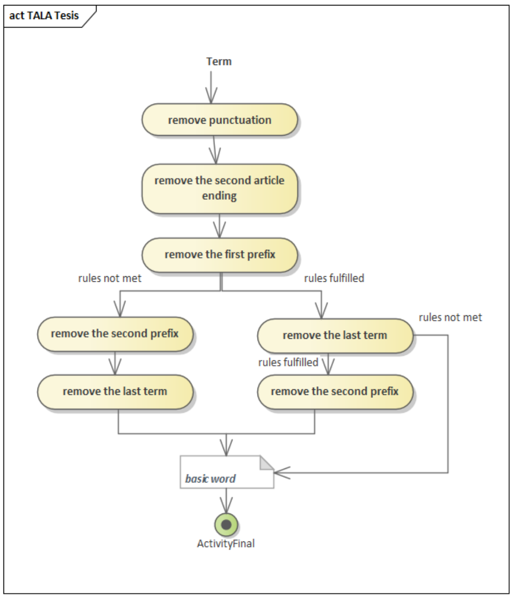
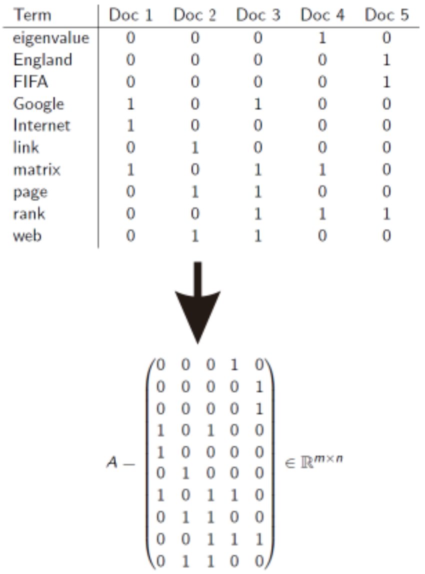


**Figure 2** Tala stemming algorithm flowchart.

Document representation techniques are used to help simplify the data calculation process. One of the document representation techniques used by the author to support the research is matrix methods according to the reference of Lars Elden and Berkant Savas, Department of Mathematics, Linkoping University, Sweden in 2012 on Data Mining using Matrix Methods[9]. This technique will outline how the news documents used will be represented in a matrix format following existing rules. The size of the matrix A, where for size *m* is determined by the number of essential words stored, then the size of the matrix A for size *n* will be determined by how many total documents are stored. Then the determination of the number 1 in the matrix if the document contains certain words. otherwise, it is 0 when it does not contain a certain word according to all essential word lines in the sample document so that it will produce a final matrix $A^{m\times n}$ representation of news and word documents Importance of the papers used. In this research process, using 100 sample news documents, the final matrix size $A^{3760\times 100}$ was produced. An overview of the process and results with this technique can be seen in Figure 3**.**



| Term | Doc 1 | Doc 2 | Doc 3 | Doc 4 | Doc 5 |
|---|---|---|---|---|---|
| eigenvalue | 0 | 0 | 0 | 1 | 0 |
| England | 0 | 0 | 0 | 0 | 1 |
| FIFA | 0 | 0 | 0 | 0 | 1 |
| Google | 1 | 0 | 1 | 0 | 0 |
| Internet | 1 | 0 | 0 | 0 | 0 |
| link | 0 | 1 | 0 | 0 | 0 |
| matrix | 1 | 0 | 1 | 1 | 0 |
| page | 0 | 1 | 1 | 0 | 0 |
| rank | 0 | 0 | 1 | 1 | 1 |
| web | 0 | 1 | 1 | 0 | 0 |

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{m\times n}$$

**Figure 3** Overview of the process and results of matrix methods [9]**.**

Making the matrix is a determination for the vector identification process for each sample document; after generating the matrix, we can get each of the document vectors from the resulting final matrix column. For example, an example based on Figure 3, vector doc 1, is $\vec{V}$ = (0,0,0,1,1,0,1,0,0,0) *etc*.

The technique of calculating the similarity between two objects can vary. One of the calculation techniques used by the author to calculate the similarity of the document with the search data is using Cosine Measure. The formula used Cosine Measure is [6]:

$$Cos\ \alpha = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i\, B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\ \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Information:

A = Vector A, which will be compared the similarity

B = Vector B, which will be compared the similarity

A • B = dot product between vector A and vector B

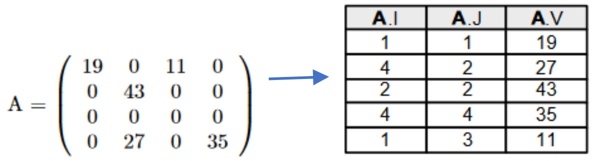
|A| = length of vector A

|B| = length of vector B

|A||B| = cross product between |A| and |B|

A sparse matrix is most of the value of is 0 zero. One way to detect that the matrix is sparse is usually determined from its sparsity value. A matrix is said to be a sparse matrix if the value of its sparsity is > 0.5. Recently, several techniques have emerged to store the sparse matrix; one of the sparse matrix storage techniques is an array representation(triples). The author will use this technique to represent the document in a matrix, as shown in Figure 3. The author chose the array representation(triples) technique because this technique is the most popular and easy to implement. This technique is also expected to reduce the matrix storage media in large quantities.

The simplest way to represent a sparse matrix is the triples (or coordinates) format. For each A(i,j) ≠ 0, the triple (i,j,A(i,j)) is stored in memory. Each entry in the triple is usually stored in a different array, and the whole matrix A is represented as three arrays A.I (row indices), A.J (column indices), and A.V (numerical values)[10], as illustrated in Figure 4**.**



$$A = \begin{pmatrix} 19 & 0 & 11 & 0 \\ 0 & 43 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 27 & 0 & 35 \end{pmatrix}$$

| A.I | A.J | A.V |
|---|---|---|
| 1 | 1 | 19 |
| 4 | 2 | 27 |
| 2 | 2 | 43 |
| 4 | 4 | 35 |
| 1 | 3 | 11 |

**Figure 4** Triples representation [10].

The author uses the concept of this technique to represent the term document in the database so that the vector representation stored in the database will be much more efficient. The following is a picture of a *database relation* developed using the array representation(triples) concept (Figure 5).

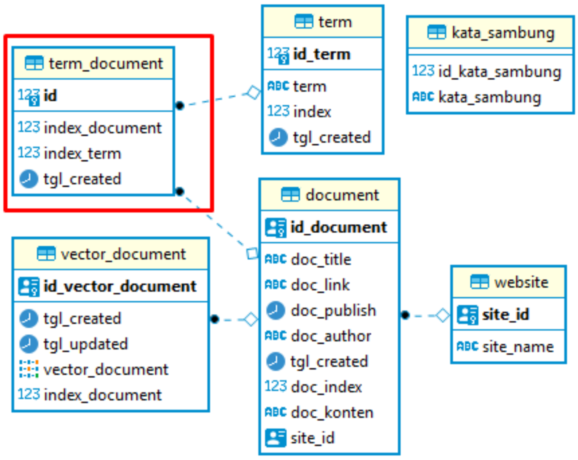


**Figure 5** Representation of document terms using array representation (triples) concept.

## 2. METHODS

This research was conducted in several stages. The stages in this study are depicted in Figure 6.

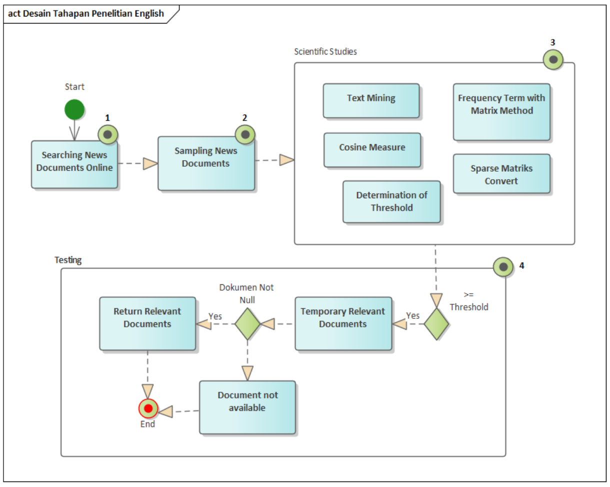

act Desain Tahapan Penelitian English
Start
1
Searching News Documents Online
2
Sampling News Documents
Scientific Studies
3
Text Mining
Frequency Term with Matrix Method
Cosine Measure
Sparse Matriks Convert
Determination of Threshold
Testing
4
>= Threshold
Yes
Temporary Relevant Documents
Dokumen Not Null
Yes
Return Relevant Documents
Document not available
End

**Figure 6** Research stages.

Figure 6 illustrates the stages of research that The author will carry out. The initial step is to find the need for news documents used as samples in this study. The sample news documents refer to news sites in Indonesia, namely. Tribunnews.com, detik.com, and coverage.com. The researcher got the document from previous research. So far, The author has obtained 210,997 documents and used experimental sample data of 10 news documents for manual calculations. Step number 2 performs a random sampling of 100 news documents from the time of publication of the latest news.

The next step, number 3, is that the researcher conducts a literature study related to the need to find relevant documents, such as implementing text mining with tala, implementing a cosine measure, conditions for determining the threshold value, and others. After step number 3 is complete, the researcher will continue testing whether it got relevant documents from search keywords.

## 3. RESULT

This chapter discusses the results of research that has been carried out.

### *3.1. Determination of the Search Document Threshold Value*

Determination of the minimum value of the cosine similarity of a document with a query is required for the requirements of a document with a certain similarity value. The author used a sample of 5 trials by testing the similarity results between news documents and queries using the Cosine Measure formula. Each test uses a different query and sample document to maximize the final results obtained. The absolute value of minimum similarity is considered as a minimum requirement to say that the news document is similar to the query. This process must be carried out first to determine the document threshold value returned to the user. The following is the final value of the resulting similarity value.

$$(threshold) = \frac{\begin{pmatrix} 0{,}155\ + 0{,}196 + 0{,}137 + 0{,}201 \\ +0{,}123 \end{pmatrix}}{5}$$

$$= \frac{0{,}812}{5} = 0{,}162$$

### *3.2. Data Illustration*

The approach found to find relevant news documents is divided into several parts, namely the first representation of news documents into an $A^{m \times n}$ Query processing, then calculating the similarity of the query representation vector to the representation matrix of the news document. Then finally, determine the cosine measure threshold value to return relevant documents by selecting the search document threshold value to produce relevant documents and get all collection documents that have similarities with other news content.

The first process is the representation of news documents into a matrix of size $A^{m \times n}$,. This process is used to accommodate some essential words that exist and which documents contain these words, and then we enter them into the $A^{m \times n}$ matrix format. The results of matrix representation can be seen in Figure 3.

The author applies the array representation(triples) concept in the term_document table contained in Figure 3. The information in the term_document table will speed up the computation and storage process because it only stores term data that only exists in certain documents. So that when it is represented in the form of a vector document, the information in the term_document table will be translated into a vector which will later contain a value of 0 or 1, which means that it does not include certain words or contains specific terms.

The following process is query processing which begins by performing a Split query in the document search process. An illustration of query processing can be seen in **Figure 7.**

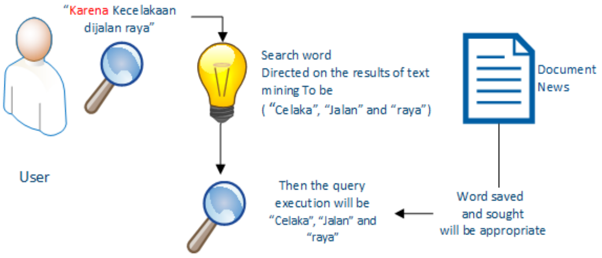

"Karena Kecelakaan dijalan raya"
User
Search word
Directed on the results of text mining To be
( "Celaka", "Jalan" and "raya")
Document News
Then the query execution will be "Celaka", "Jalan" and "raya"
Word saved and sought will be appropriate

**Figure 7** Illustration of word processing using text mining techniques.

Based on Figure 7, it can be concluded that this process overcome the word meaning errors according to the original word or its constituents. Words stored in the database or text mining results only keep essential words or words that are not conjunctions. The author does not use conjunctions in Indonesian because all documents will contain these conjunctions to eliminate the unique identity of each news document. When the user enters a word with a junction, the word will be destroyed. Therefore, it is necessary to process a split query at the beginning of the process to direct the search words that match the words stored in the database.

The following process calculates the matrix representation of news documents with vector representation of the query. The author uses the cosine measure technique to find similarities between them. The similarity threshold value of news documents was found to be 0.162. This means that when returning relevant documents, the condition for the cosine measure value between the query representation vector and the news document representation matrix must be greater than or equal to >= 0.162. Therefore, the documents provided must meet the threshold value requirements, so it is hoped that the results of the returned records are truly relevant or by the search data. This test has been carried out by the author and shows promising results with the average execution time of the function in the database for 100 documents with an average of 200 - 210 words contained in the document. It only takes 5-10 seconds.

## 4. CONCLUSION

Based on the research that has been done, it can be approaches and results that,

1. Based on the results of testing the determination of the minimum cosine limit that has been carried out, the minimum cosine similarity value required to obtain the relevant documents is 0,162. This means that when displaying relevant news documents, a threshold value is needed as a reference value to be displayed based on the similarity between certain documents and the entered query. So that when displaying the relevant document, it must have a cosine value >= threshold (0,162).

2. Using matrix methods becomes efficient to calculate the similarity with the cosine measure and combine the sparse matrix storage technique with an array representation(triples). It will speed up the computing process and reduce storage when the implementation process is carried out in computer programs. This test has been carried out by the author and shows promising results with the average execution time of the function in the database for 100 documents with an average of 200 - 210 words contained in the document. It only takes 5-10 seconds.

## ACKNOWLEDGMENT

The author would like to thank all those who have helped and facilitated this research process.

## REFERENCES

[1] Asosiasi Penyelenggara Jasa Internet Indonesia, Laporan Survei Internet APJII 2019 – 2020, Asos. Penyelenggara Jasa Internet Indones., vol. 2020, pp. 1–146, 2020, [Online]. Available: https://apjii.or.id/survei.

[2] Katadata Insight Center, Masyarakat Paling Banyak Mengakses Informasi dari Media Sosial, Dkadata.co.id, p. 670, 2020, [Online]. Available: https://databoks.katadata.co.id/datapublish/2020/11/23/masyarakat-paling-banyak-mengakses-informasi-dari-media-sosial.

[3] Alexa - Top Sites in Indonesia - Alexa. https://www.alexa.com/topsites/countries/ID (accessed Dec. 17, 2021).

[4] M. A. Rohim, Implementasi Ekstraksi Web (Web Scraping) Pada Situs Berita Menggunakan Metode Ekspresi Reguler, Digital Repository, pp. 68–74, 2018.

[5] Y. D. Hadiyat, Clickbait on Indonesia Online Media, J. Pekommas, vol. 4, no. 1, p. 1, 2019, doi: 10.30818/jpkm.2019.2040101.

[6] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi, J. Tek. Elektro, vol. 9, no. 1, pp. 18–23, 2017, doi: 10.15294/jte.v9i1.10955.

[7] IR-NLP Lab - CSUI. https://ir.cs.ui.ac.id/new/ (accessed Dec. 17, 2021).

[8] M. S. H. Simarangkir, Studi Perbandingan Algoritma - Algoritma Stemming Untuk Dokumen Teks Bahasa Indonesia, J. Inkofar, vol. 1, no. 1, pp. 40–46, 2017, doi: 10.46846/jurnalinkofar.v1i1.2.

[9] L. Eld, TANA07 : Data Mining using Matrix Methods, 2012.

[10] A. Buluç, J. Gilbert, and V. B. Shah, Implementing Sparse Matrices for Graph Algorithms, Graph Algorithms Lang. Linear Algebr., vol. 94720, pp. 287–313, 2011, doi: 10.1137/1.9780898719918.ch13.