Journal of Telematics and Informatics (JTI)
Vol.5, No.2, September 2017, pp. 56~61
ISSN: 2303-3703

Cosine Similarity Measurement for Indonesian Publication Recommender System

Darso, Imam Much Ibnu Subroto, Sri Arttini Dwi Prasetyowati
Department of Informatics Engineering
Amikom Purwokerto
e-mail: darso@amikompurwokerto.ac.id

Abstract

The number of publications is increasing every time and causing the flood of articles so often the relevant knowledge and information can not be delivered properly. As a public scientific site, the IPI (Index of Indonesian publication) garuda portal has a publication collection of over 4000 journals in Indonesia in a managed database. The system has the potential to provide benefits in the form of information and knowledge for Indonesian users in particular and researchers around the world in general. Most of the papers are written in Indonesian as the official language of the country so this paper offers a model and implementation of an article recommendation system that has a high relevance to the paper being read by the user. Recommendations are made by calculating the similarity between related documents using the cosine equation approach. Natural language processing is necessary to process the data from the original text into terms that are based on the calculation of rank-based similarity. Preprosecessing includes text clearance, tokenization, stop word filtering, and stemming. Normalization of the TF-IDF term frequency is used as a weighted vector for cosine calculation. The experimental results in the study showed a good result that is recall 0.96 and precision 0.66

Keywords: *Information, Recommendation System, Cosine Similarity, IPI*

1. Introduction

The number of articles and items offered by the article search provided by the website to the user causes the overflow of information that makes users often have difficulty in finding the information that is really sought or needed. In search of articles in the journal publications of website providers, for example, the difficulties faced by users as prospective readers is when looking for articles that fit and according to taste or preferences. Readers also want from indexers to get relevant and quality articles easily and accurately [1].

Most of the current journalist sites have large and complex file structures to make the information required by potential readers unable to be delivered completely. There are not many systems on the website that provide ease and speed to users in finding and getting relevant article information that is being searched. Therefore, for journal websites, a system is required to adapt to each user's profile and may advise users on relevant articles. This personalized website system is called a recommender system (RS). RS works to help users find articles by providing information based on user's level of taste. The user's sense can be known from some data, for example from visits and the length of time a user accesses a web page, the user searches by writing a query or keyword in the search menu, or from the user's download data in the past and the number of access by showing the rating of other users.

The Garuda Portal of Indonesian Publication Index (IPI) is designed for browsing, indexing, abstracting, monitoring and upgrading of scientific publication standards in Indonesia. This site was established in 2012 initiated by IAES (Institute of Advanced Engineering and Science Indonesia Section). As a scientific publication website, the Garuda portal IPI has more than 4000 journals in Indonesia in the database. So much publicity is increasing every year on this site that the overflow of data so that the function of the site cannot be delivered completely to the visitors. Therefore garuda portal requires a recommender system of article search so as to facilitate the process of visitor search.

In this study, the authors aim to make the design of recommender system of article search in Indonesian by using cosine similarity method so as to know the performance of recommendation system using precision and recall. Then make the design of the application by using an efficient algorithm by using the cosine similarity method to calculate the similarity between documents, making it easier for users to get the desired article information.

2. Research Method

The purpose of this research is to design a recommendation system on the website of IPI (Indonesia Publication Index) portal gallery portal using Cosine Similarity method.

In another study, the use of dataset obtained binary data reader to know every article that was often read then Cosine Similarity algorithm was a method to predict the reference of reader interest to news articles. The algorithm of this system used Collaborative Filtering by entering the amount of past data or history of articles ever read. [2] Researchers then used the algorithm which only worked optimally when handling documents in English. In this study on the dependence of the search engine API, whereas the API has a usage restriction for each day [3]. Another research was recommendation system that could provide book collections in the library that could be utilized by the user [4].

2.1. Recomender System

Recommendation Systems are software and techniques that provide recommendations for use by users. Recommendations given are related to the decision-making process. Netflix, online movie borrowing, Amazon.com which uses a recommendation system that helps users select a book, are some of the many sites that apply recommendation systems to users who have different options [5].

2.2. Cosine Similarity

Cosine Similarity is a method used to calculate the similarity (level of similarity) between two objects. The main advantage of the cosine similarity method is not affected by the short length of a document [6]. In general the calculation of this method is based on the vector space similarity measure. This cosine similarity method calculates the similarity between two objects (eg D1 and D2) expressed in two vectors by using keywords (keywords) of a document as a measure. This method of measurement of conformity has several advantages, namely the normalization of the length of the document. This minimizes the effect of document length.

$$\cos(\emptyset) = \frac{A \,.\, B}{||A|| \,.\, ||B||} \qquad (1)$$

Where :

A ·B = vectors dot product of A and B are computed with $\sum A.B$

And

||A|| = The length of the vector A, calculated by $\sqrt{\sum A^2}$

||B|| = The length of the vector B, calculated by $\sqrt{\sum B^2}$

Then it can be formulated as follows :

$$\cos(\emptyset) = \frac{\mathrm{A}\,.\mathrm{B}}{||\mathrm{A}||.||\mathrm{B}||} = \frac{\sum \mathrm{A}\,.\mathrm{B}}{\sqrt{\sum \mathrm{A}^2 \sum \mathrm{B}^2}} \qquad (2)$$

2.3. Text Mining

Text mining has the definition of data mining in the form of text where the data source is usually obtained from the document, and the goal is to find words that can represent the contents of the document so that it can be analyzed the relationship between documents.

Common stages are done in Text Mining as follows

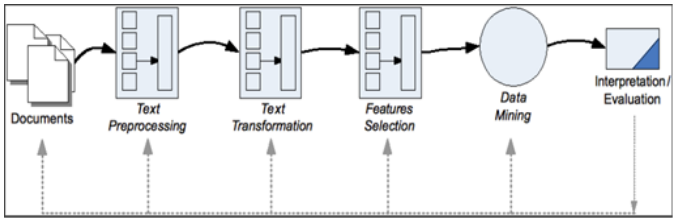

Figure 1. Stages of Text Mining [7]

Text mining process includes tokenizing, filtering, stemming, and tagging.

2.4. Normalization

The global weight of a term I in the inverse document frequency (Idfi) approach can be formulated as follows:

$$Idfi = log^2 (N/D_{ji}) \quad (3)$$

Where N denotes the number of articles in the document collection, *dji* is the frequency document from term *i*, and is equal to the number of documents containing term *i*. Log^2 is used to minimize its effect relative to *tfij*.
The weight of term *I* is calculated using *tf-idf* defined as follows.

$$Wij = tfij \times idfi \quad (4)$$

2.5. Vector Space Model

Similarity measures of vector space models are used to rank information retrieval documents [8]. An example of a three-dimensional vector space model for two D1 and D2 documents, one query user Q1, and three terms T1, T2 and T3 are shown in the following figure.

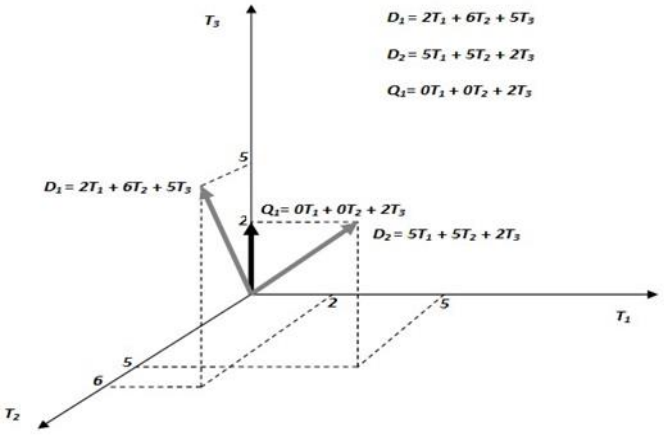

Figure 2. Sample Vector Space Model D1, D2, Q1 [7]

2.6. Evaluation Method

The final component of the recommendation system is relevancy assessment. "Recall thus relates to the ability of the system to retrieve relevant documents" [9]. The opinion can be interpreted that recall (acquisition) was associated with the ability of a system in finding the relevant documents. This means that the recall was part of the information retrieval process that can be used as a measure of the effectiveness of an information retrieval system.

"Recall corresponds to the ability of the system to call the relevant document, whereas precision is related to the system's ability not to call irrelevant documents", [10].

$$Precesion = \frac{Number\ of\ Relevant\ Documents\ found}{Number\ of\ all\ Documents\ found}\ x\ 100\ \% \quad (5)$$

$$Recall = \frac{Number\ of\ Relevant\ Documents\ found}{The\ number\ of\ all\ relevant\ Documents\ in\ the\ collection}\ x\ 100\ \% \quad (6)$$

The above sizes are usually rated in percentage form, 1 to 100%. An information system will be considered good if the level of recall and precision is high.

3. Results and Analysis

3.1. Similarity Measurement

The process of counting the similarity in the flow diagram as in Figure 3.5 begins to take the keyword (query) from the user, and then apply *stopword removal* and stemming so that keyword is resulted but can represent the query. Based on this keyword, the system looks into the cache. If in the system already exists the keyword then the system instantly sort documents similar to the keyword will return it to the user. If the keyword is not contained in the cache, meaning there has never been any similarity calculation based on that keyword, the system calculates the similarity between the keywords with the list of documents represented by the terms in the index. The results of these calculations are stored into the cache table. Furthermore the system displays the recommendation of the document title and sort by the highest value based on the similarity of keywords. Here is a flow chart of the similarity process.

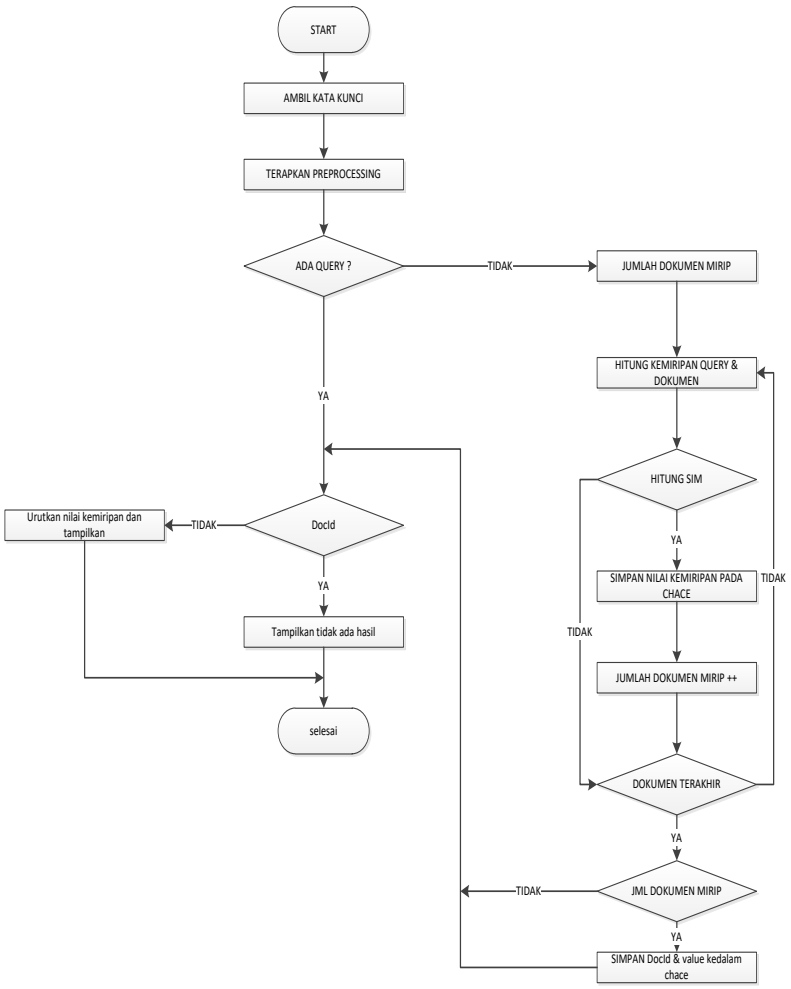

Figure 3. Flow chart calculates the similarity

3.2. Search Interface

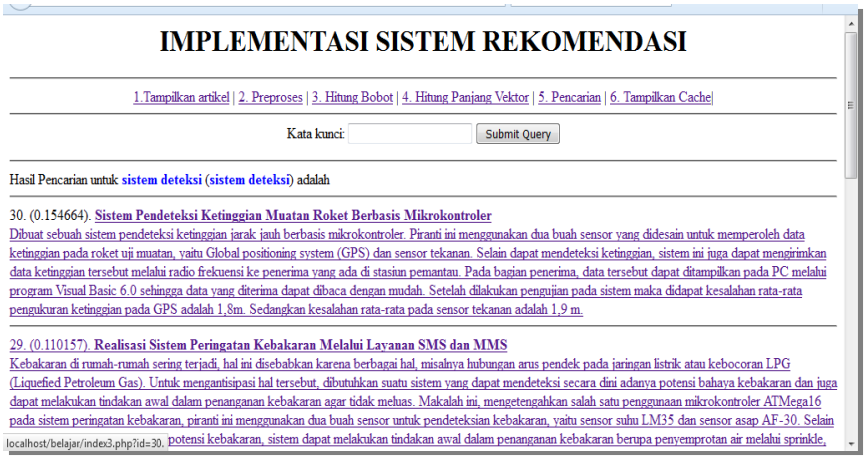

Figure 4. Search Simulation

In the Search menu process function displays articles based on keywords entered in the search field by the user. From these keywords the system will calculate the similarity between documents in the database.

3.3. Recommended System Interface

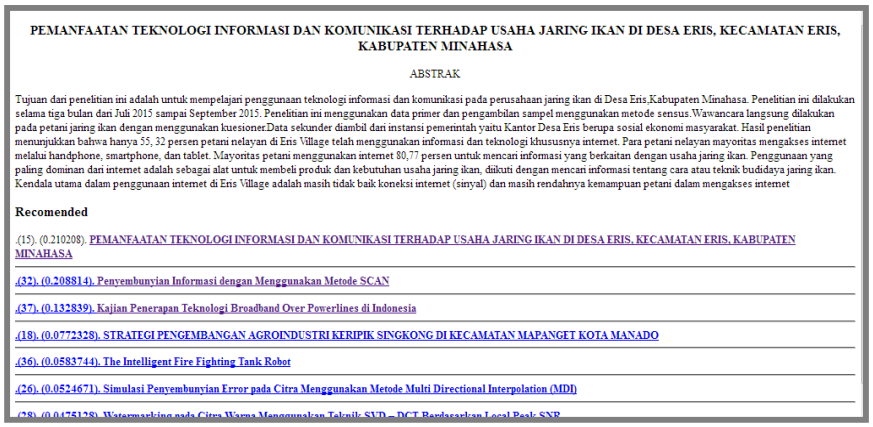

Figure 5. Recommended System Page

This recommendation system page appears when the user has done a search and then gets the title of the article of interest. This page will display the title of the article, abstract and recommend the relevant article title according to the user's interest then sort by the highest level of calculation from the system.

3.4. System Evaluation

This stage aims to test the ability of RS system (Recommender System) by measuring the quality of retrieval that is doing precision and recall testing.

The performance of the recommendation system on keyword samples 1 to 5 can be seen in Table 1.The performance of the system is good enough because the average average precision is about 0.664 which means the relevant documents are returned successfully. The average of each recall point, 0.962 in good system performance means the corresponding document successfully returned relevant to the given query.

Table 1. Average Result of Recall and Precision Test

Query	Recall	Precision
Pengolahan citra digital	1	0.45
Ekonomi pertanian	1	0.92
Teknologi Informasi	1	0.8
Sistem digital	0.81	0.75
potensi hasil usaha tani	1	0.4
Average	0.962	0.664

Based on the research that has been done, the relevancy of an article sought was determined by the user himself. The system only returned documents related to the query entered by the user.

4. Conclusion

The performance of recommendation system is quite good because with the average value of precision about 0.664 from the relevant documents found, then the average value of recall point 0.962 of the documents returned successfully relevant to the given query.

References

[1] I. M. I. Subroto, T. Sutikno, and D. Stiawan, "The Architecture of Indonesian Publication Index: A Major Indonesian Academic Database," *TELKOMNIKA (Telecommunication Computing Electronics and Control),* vol. 12, p. 1, 2014.

[2] a. Rajendra LVN, "Recommending News Articles using Cosine Similarity Function," 2014.

[3] I. K. R. d. B. Zaman, "Sistem Rekomendasi Pustaka Dengan Metode Automatic Query Expansion," 2013.

[4] Adi Wibowo, "Recommender System di Perpustakaan Universitas Kristen Petra menggunakan Rocchio Relevance Feedback dan Cosine Similarity," 2014.

[5] F. R. e. al., "Recommender Systems Handbook," 2011.

[6] I. R. d. S. Rozas, R., "Sistem Pemilihan Kontrol Keamanan Informasi Berbasis ISO 27001, ," in *Seminar Nasional Pascasarjana XI*, ITS, 2011.

[7] M. C. Harlian. (2011). *Text Mining*. Available: lecturer.eepis-ts.edu/~iwanarif/kuliah/dm/6Text%20Mining.pdf

[8] s. G, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. New York: Addison-Wesley, 1889.

[9] G. Chowdhury, G, "Introduction to Modrern Information Retrieval," ed. London: Library Association Publishing, 1999.

[10] J. Hasugian, "Penelusuran Informasi Ilmiah Secara Online: Perlakuan terhadap Seorang Pencari Informasi Sebagai Real User". Dalam Pustaha: Jurnal Studi Perpustakaan dan Informasi " vol. 2, 2006.