Scientific Journal of Informatics
Vol. 4, No. 2, November 2017

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Information Retrieval System for Determining The Title of Journal Trends in Indonesian Language Using TF-IDF and Naïve Bayes Classifier

Wandha Budhi Trihanto[1], Riza Arifudin[2], Much Aziz Muslim[3]

[1,2,3]Computer Science Department, Mathematics and Natural Sciences Faculty, Semarang State University
Email : [1]wandhabudhi@gmail.com, [2]rizaarifudin@mail.unnes.ac.id, [3]a212muslim@yahoo.com

Abstract

The journal is known as one of the relevant serial literature that can support a researcher in doing his research. In it's development journal has two formats that can be accessed by library users namely: printed format and digital format. Then from the number of published journals, not accompanied by the growing amount of information and knowledge that can be retrieved from these documents. The TF-IDF method is one of the fastest and most efficient text mining methods to extract useful words as the value of information from a document. This method combines two concepts of weight calculation that is the frequency of word appearance on a particular document and the inverse frequency of documents containing the word. Furthermore, data analysis of journal title is done by Naïve Bayes Classifier method. The purpose of the research is to build a website-based information retrieval system that can help to classify and define trends from Indonesian journal titles. This research produces a system that can be used to classify journal titles in Indonesian language, with system accuracy in determining the classification of 90,6% and 9,4% error rate. The highest percentage result that became the trend of title classification was decision support system category which was 24.7%.

Keywords*:* TF-IDF, Naïve Bayes Classifier, Trends.

1. INTRODUCTION

The journal is known as one of the relevant serial literature that can support a researcher in doing his research. In its development the journal has two formats that can be accessed by the user (library user) that is: printed format and digital format [1]. Of the many research results in the form of journals published online, not accompanied by the growing amount of information and knowledge that can be taken from these documents.

Text mining in general is the theory of processing a large collection of documents that exist from time to time by using some analysis, the purpose of text processing is to know and extract useful information from data sources with the identification and exploration of interesting patterns in the case of text mining, The data used is a collection or collection of unstructured documents and requires the existence of a grouping for known similar information [2].

In this research we will use TF-IDF and Naïve Bayes Classifier method to determine the trend of journal titles in Indonesia. The TF-IDF (Term Frequency Inverse Document

Frequency) method is one of the fastest and most efficient text mining methods to extract useful words as the value of information from a document [3]. The TF-IDF (Term Frequency Inverse Document Frequency) method is a way to weight the relationship of a term to a document. This method combines two concepts for weight calculation, ie the frequency of occurrence of a word within a particular document and the inverse frequency of the document containing the word. The frequency of occurrences of words in a given document indicates how important the word is in the document [4].

Furthermore, the final stage of this research data mining analysis is done by using Naïve Bayes Classifier method. Naïve Bayes is one of the meode on probabilistic reasoning. Naïve Bayes algorithm aims to classify data in a particular class, then the pattern is the journal title class that is becoming a trend. The advantages of the Naïve Bayes Classifier method are simple but have high accuracy. Based on the experimental results [5], the Naïve Bayes Classifier on previous research proved to be used effectively for the categorization of Indonesian text with an accuracy of 90%. The simple Naïve Bayes Classifier algorithm and its high speed in the training and classification process make this algorithm appealing to be used as one of the classification methods. The purpose of the research is to build a website-based information retrieval system that can help to classify and define trends from Indonesian journal titles. In addition, this study also aims to apply the TF-IDF and Naïve Bayes Classifier methods in determining the trend of Indonesian journal titles.

2. METHODS

2.1 Information Retrieval System

The information retrieval system for determining the trend of Indonesian journal titles uses a waterfall system development model, where it proposes an approach to systematic and sequential software development that begins at the level and progress of the system throughout analysis, design, code and testing [6]. Information retrieval is divided into the following sections [7]:

1. Text Operations, including selection of words in the query or document (term selection) in the process of transforming the document or query into term index (index of words).
2. Query formulation, giving weight to the query index of words.
3. Ranking, searching for documents relevant to the query and undoing the documents based on their compatibility with the query.
4. Indexing, builds the index data base of the document collection First done before the document search is done

2.2 Text Mining

Text mining or text analytics is a term that describes a technology capable of analyzing semi-structured and unstructured text data, this is what distinguishes it from data mining where data mining processes data that are structured. In essence, text mining is an interdisciplinary field that refers to information retrieval, data mining, machine

learning, statistics, and linguistic computation [8]. In general, the concept of text mining work is similar to data mining, which is predictive digging and descriptive digging. Text mining extracts information from the text and converts it into a meaningful numerical index [9]. In order to obtain the final goal of text mining, it takes several stages of the process to be performed as shown in Figure 1 [10]. The selected data to be analyzed first will pass through the Pre-process stage and text representation, until finally knowledge discovery can be performed.

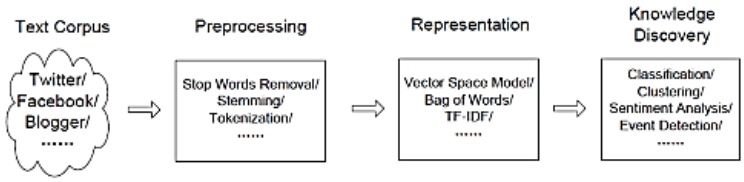

Figure 1. Text process analysis framework in text mining

2.3 TF-IDF Method

Needs analysis in this discussion consists of two sub-discussion of discussion of TF-IDF method and Naïve Bayes Classifier method. The TF-IDF method is used to determine the weight value of each word (term) contained in each document. The TF-IDF equation scheme is represented by the following equation [11]:

$$w_{m,i} = freq_{m,i} \times log\left(\frac{N}{nm}\right) \quad (1)$$

2.4 Naïve Bayes Classifier

The Naïve Bayes Classifier procedure there are 2 stages in the process of text classification. The first stage is training on the set of journal titles (data training). While the second stage is the process of classification of documents that have not known the category (data testing). The Naïve Bayes method will calculate the probability of each case. The value of the target attribute in each sample data. Then the Naïve Bayes Classifier will classify the sample data to the class that has the highest probability value [12]. The formula of the Naïve Bayes Classifier method is as follows [13]:

$$P(H|X) = \frac{P(X|H)P(H)}{PX} \quad (2)$$

3. RESULT AND DISCUSSION

3.1 Preprocessing

Before the classification process on the journal title data, the data will enter the tokenisasi stage to make the title a word per word, then from each word will be removed stopword it. Stopword is found by counting the number of occurrences of words. Then, the number of word occurrences is sorted from the most words appear, words that appear big will be the candidate stopword. The example of the found stopword is shown in Table 1.

Table 1. *Stopword Data*

Words
Yang
Dan
Untuk
Dalam
Dari
Ke
Di

In addition to using the number of occurrences of large words, the selected stopword is a foreground like: “dalam”, “di”, “ke”, “dari”, connecting words like: "the", "and", and using subjective options According to the needs of the journal title data. The entire list of used stopwords is in Appendix 2. After all the stopwords found in the title are removed . Then the next process is stemming, that is change the word into the word base. An example of the stemming results in this study is shown in Table 2.

Table 2. Results stemming title

Title	Stemming result
Audit Sistem Informasi Aplikasi Penjualan Tunai pada PT AJ	sistem informasi jual tunai
Analisa Sistem Informasi Penjualan Online *E-Commerce* pada CV Selaras Batik Deskriptif	sistem informasi *e-commerse*
Rancang Bangun Sistem Pakar Berbasis Web Mendiagnosa Penyakit pada Sapi Perah	sistem pakar web diagnosa penyakit
Sistem Informasi Geografis Pengelolaan Monitoring Persebaran Aset Wakaf	sistem informasi geografis
Penerapan *Fuzzy Inference System* Takagi Sugeno Kang pada Sistem Pakar Diagnosa Penyakit Gigi	sistem pakar diagnosa penyakit

3.2 Weighted Phase of TF-IDF

After the process of changing the word base, then done the weighting. The first process is to calculate the occurrence of words or term frequency (TF) for each document based on words that have been used as a basic word on the stemming process. Suppose for the word "sistem pakar" appears in the title 1 as much as 1 time then the column table title 1 is written number 1, then because in the title 3 not found the word "sistem pakar" then in the column title 3 is written number 0. The example of the result of the word (TF) can be seen in Table 3.

Table 3. Word Appearance (TF)

	Title 1	Title 2	Title 3	Title 4	Title 5
Sistem pakar	1	1	0	0	0
Web	1	0	0	0	0
Diagnosa	1	1	0	0	0
Penyakit	1	1	0	0	0
Fuzzy	0	1	0	0	0
Takagi	0	1	0	0	0
Sugeno	0	1	0	0	0
Kang	0	1	0	0	0
Sistem pendukung keputusan	0	0	1	1	1
Promethee	0	0	1	0	0
Analytical hierarcy process	0	0	0	1	0
Electre	0	0	0	0	1

The next process is to calculate the inverse document frequency (IDF) which is the normalization of word frequency. To calculate the IDF, is by the formula:

$$IDF = log(\frac{nm}{n}) \quad (3)$$

Example to calculate the IDF of the word "sistem pakar" whose occurrence (nm) in 2 titles and the amount of title data (n) is 5, then it is obtained:

$$IDF\ (sistem\ pakar) = log(\frac{nm}{n}) = log(\frac{2}{5}) = \ 0{,}39794$$

The next step is to calculate the TF-IDF value for each document generated from the number of word occurrence frequency ($freq_{m,i}$) multiplied by the IDF value, To calculate TF-IDF, is by the formula:

$$w_{m,i} = freq_{m,i} \times log\left(\frac{N}{nm}\right) \quad (4)$$

Calculates the weight of the word "sistem pakar" in heading 1 by the number of times that the word appears $freq_{m,i} = 0{,}4$ dan IDF = 0,39794

$$w_{(sistem\ pakar, judul1)} = 0{,}4 \times log\left(\frac{5}{2}\right) = 0{,}159176$$

Then the weight of the word "sistem pakar" in title 1 is 0.159176. For table weighted results using TF-IDF is shown in Table 4.

Table 4. TF-IDF Weighted Results

	TF*IDF				
Term	Title 1	Title 2	Title 3	Title 4	Title 5
sistem pakar	0,159176	0,159176	0	0	0
Web	0,139794	0	0	0	0
Diagnosa	0,159176	0,159176	0	0	0
Penyakit	0,159176	0,159176	0	0	0
Fuzzy	0	0,139794	0	0	0
Takagi	0	0,139794	0	0	0
Sugeno	0	0,139794	0	0	0
Kang	0	0,139794	0	0	0
sistem pendukung keputusan	0	0	0,133109	0,133109	0,133109
promethee	0	0	0,139794	0	0
analytical hierarchy process	0	0	0	0,139794	0
Elecre	0	0	0	0	0,139794

3.3 Probability Calculation Phase

The Naïve Bayes method uses probability theory as the theoretical basis. The probability calculation is done to get the result of value which will lead the document into the class or category. There are two stages in the text classification process. The first stage is training on the set of journal titles (data training). While the second stage is the process of classification of documents that have not known the category (data testing).

Where f (c_i) is the value of the appearance of the wkj feature, | W | Is the number of words/features used. While f (w_{kj}, ci) is the number of occurrences of the word wkj in category vi. The number of words in each class is expressed as n. The following is the application of the naïve bayes classifier method in determining how the process of category determination. Starting from determining the data title of the journal to be used as training data then determine the category of each data manually. Table 5 is the title data training that has been specified category.

Tabel 5. Data Training Titles Already Have Category

Title	Category	Word appearance
Rancang Bangun Sistem Pakar Berbasis Web Untuk Mendiagnosa Penyakit pada Sapi Perah	Sistem Pakar	Sistem Pakar(1), Diagnosa(1), Penyakit(1)
Penerapan Fuzzy Inference System Takagi-Sugeno-Kang pada Sistem Pakar Diagnosa Penyakit Gigi	Sistem Pakar	Sistem Pakar(1), Diagnosa(1), Penyakit(1)
Sistem Pendukung Keputusan Pemilihan Gadget Android Menggunakan Metode Promethee	Sistem Pendukung Keputusan	Sistem Pendukung Keputusan(1)
Rancangan Sistem Pendukung Keputusan Peminatan Jenjang dan Jurusan dengan Menggunakan Metode Analytical Hierarchy Process (Studi Kasus pada Siswa SMP Negeri 39	Sistem Pendukung Keputusan	Sistem Pendukung Keputusan(1), AHP(1)
Implementasi Metode Electre pada Sistem Pendukung Keputusan SNMPTN Jalur Undangan	Sistem Pendukung Keputusan	Sistem Pendukung Keputusan(1)

Training data and categories have been obtained then start testing data testing that has not been known category. Data testing of unknown title categories can be seen in Table 6. The data testing is what the categorie will automatically determine to use.

Table 6. Data Testing Unknown Title Categorization

Title	Category	Word appearance
Entropybased Fuzzy AHP Sebagai Pendukung Keputusan Penempatan Bidan Kota Banjar Baru	?	AHP(1), Sistem Pendukung Keputusan(1)
Sistem Pakar Diagnosis Penyakit Unggas dengan Metode *Certainty Factor*	?	Sistem Pakar(1), Penyakit(1)

Next count the number of occurrences of the word (term) in the title 1 and title 2. In title 1 only appears the word "AHP" as much as 1 time, then in title 2 there is the word "expert system" 1 time and "disease" 1 times. For keywords that do not appear in the title then it is written "0". The term appearance tables in the data testing can be seen in Table 7.

Table 7. Occurrence of Term in Data Testing

Title	Sistem Pendukung Keputusan	Sistem Pakar	Diagnosa	Penyakit	AHP
Title 1	1	0	0	0	1
Title 2	0	1	0	1	0

The next step is to look for the probability value of word appearance for each category, from training data. To determine the probability value of words that appear in each category is by the formula:

$$P\ (\mathrm{w_k} \mid \mathrm{v}_j) = \frac{f(w_{kj}, c_i)+1}{f c_i + |w|} \quad (5)$$

The results of word probability calculations for each category can be seen in Table 8.

Table 8. Word Probability Table for Each Category

Category	$P(c_i)$	$P\ (\mathrm{w_k} \mid \mathrm{v}_j)$				
		Sistem Pendukung Keputusan	Sistem Pakar	Diagnosa	Penyakit	AHP
Sistem Pakar	0,4	0,185185185	0,244444	0,244444	0,244444	0,185185
Sistem Pendukung Keputusan	0,6	0,248214286	0,178571	0,178571	0,178571	0,203571

The final step in the category determination is to use the following formula:

$$V_{MAP} = \text{argmax}_{\mathrm{vj} \in \mathrm{V}} P(\mathrm{vj})\ \Pi\ \mathrm{i}\ P(\mathrm{ai} \mid \mathrm{v}) \quad (6)$$

Specify title category 1

$P(sistem\ pendukung\ keputusan|judul\ 1)$
$= P(sitem\ pendukung\ keputusan)P(sistem\ pendukung\ keputusan\ )|AHP)$
$P(sistem\ pendukung\ keputusan|sistem\ pendukung\ keputusan)$
$= 0{,}6\ .\ 0{,}2036\ .\ 0{,}2482$
$= 0{,}0303$

$$P(sistem\ pakar|judul\ 1) = P(sistem\ pakar)P(sistem\ pakar|AHP)$$
$$= 0{,}4\ .\ 0{,}1851$$
$$= 0{,}074$$

Because *P*(sistem pendukung keputusan | judul 1) > *P*(sistem pakar | judul 1) then the category of title 1 is sistem pendukung keputusan.

- **Specify title category 2**

$$P(\text{sistem pakar}|\text{judul 2}) = P(sitem\ pakar)P(sistem\ pakar|\text{sistem pakar})$$
$$= 0{,}6\ .\ 0{,}244\ .\ 0{,}244$$
$$= 0{,}0357$$

$P(\text{sistem pendukung keputusan}|\text{judul 2})$
$= P(sistem\ pendukung\ keputusan)P(sistem\ pakar|sistem\ pendukung\ keputusan)$
$P(penyakit|sistem\ pendukung\ keputusan)$
$= 0{,}6\ .\ 0{,}1786\ .\ 0{,}1786$
$= 0{,}0191$

Because *P*(sistem pakar | judul 2) > *P*(sistem pendukung keputusan | judul 2) then the category of title 2 is sistem pakar.

3.4 Discussion

Based on the results of the application of Naïve Bayes Classifier method has been done, it can be known classification the title of the Indonesian computer science journal from the DOAJ website. Journal titles are divided into 13 classes including decision support systems, expert systems, information systems, e-commerce, e-learning, computer networks, artificial neural networks, image processing, cryptography, artificial intelligence, geographic information systems, educational games, and applications Mobile. In the classification process must go through several stages of data retrieval process, the division into training data and data testing, preprocessing stage and classification.

Journal data retrieval process is done by using API DOAJ, then the data that has been got divided into 2 that is data training and data testing. Then enter the preprocessing stage that is done tokenizing process, stopword and stemming. After going through the process of preprocessing the process of classifying this process is a prosen where the data training and data testing is calculated probability.

After getting the class or each category then do the trends counting from the classification of the title, Here is a description of the data for the calculation of trends in the system. The test of this system is only done with data training as much as 281, data testing as much as 85 data, and 13 classification categories that have been determined. The classification includes decision support systems, expert systems, information systems, e-commerce, e-learning, computer networks, artificial neural networks, image processing, cryptography, artificial intelligence, geographic information systems, mobile applications, and educational games. From the calculation scenario of this system generated trend percentage of title classification that is. sistem pendukung keputusan 24,7%, sistem informasi 21,7%, sistem pakar 14,6%, e-learning 9,8%, e-commerce 3,5%, kriptografi 6,7%, game 5,2%, kecerdasan buatan 5,3%,

sistem informasi geografis 1,8%, jaringan komputer 1,6%, pengolahan citra 2,1%, jaringan syaraf tiruan 1,6%, aplikasi mobile 1,4%. The percentage of trend classification of journal title can be seen in Figure 2.

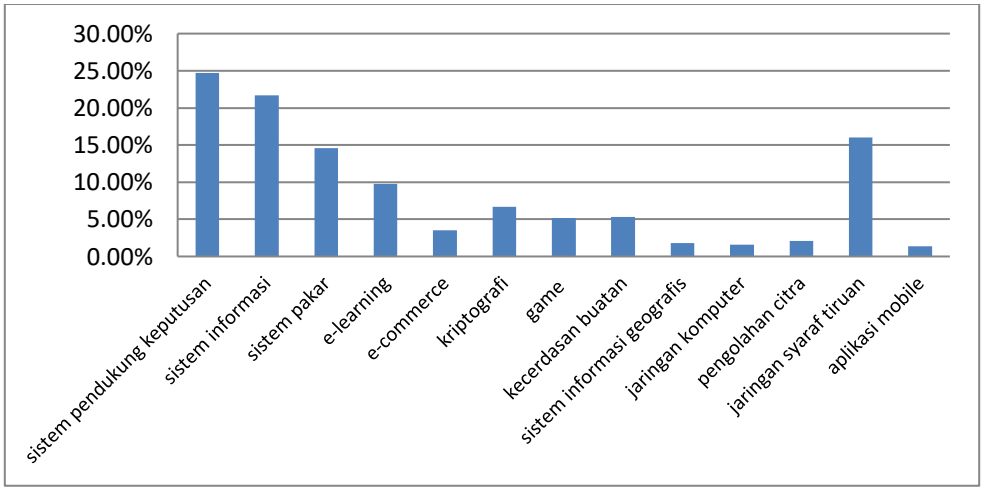

Figure 2. Percentage of Classified Trends of Journal Title Classification

After the percentage of the journal title classification is obtained, an evaluation of the classification data from the data testing is conducted. This evaluation will result in accuracy and error rate of the classification process. To calculate the value of accuracy and error rate is as follows.

$$Accuracy = \frac{Jumlah\ prediksi\ yang\ benar}{Jumlah\ prediksi\ benar\ dan\ salah} \quad (7)$$

$$= 8 + 25 + 14 + 11 + 1 + 6 + 3 + 5 + 1 + 3 + 0 \,/\, 85$$

$$= 77 \,/\, 85 = 0{,}906 = 90{,}6\%$$

$$Error\ Rate = 1 - Accuracy \quad (8)$$

$$= 1 - 0{,}906 = 0{,}094 = 9{,}4\%$$

So obtained the accuracy of 90.6%. This value is obtained from the calculation of the sum of all data that is predicted correctly divided by all the test data in the form of error rate of 9.4% obtained from the calculation of the sum of value 1 minus the result of accuracy. The advantages of this system is to use this system we can find trends from the title of the journal in the field of computer science available on DOAJ website. In this system also has a shortage of training data that is used still use labeling manually so it is difficult when labeling done on the data training in large numbers.

4. CONCLUSION

Based on the description of the results and discussion of this research, it can be concluded that the application of TF-IDF and Naïve Bayes Classifier method in the system to determine the trend of Indonesian journal titles is using several stages. First, data retrieval using DOAJ API. Second, the preprocessing stage consists of tokenization, stemming and filtering. Third, the word weighting stage with TF-IDF and the last stage is the process of classifying the test data using Naïve Bayes Classifier by calculating the probability value of each text. Then from the classification of training data and data testing will be obtained trends from the title of the journal by class or category that has the largest to smallest members. The results of the 5th grade or category with the highest number of trends are decision support system 24.7%, information systems 21.7%, expert systems 14.6%, e-learning 9.8%, cryptography 6.7%.

5. REFERENCES

[1] Rusydi, I. 2014. Pemanfaatan E-Journal sebagai Media Informasi Digital. *Jurnal Iqra'.* Vol. 8(20), pp. 200-210.

[2] Somantri, O., & Wiyono, S. 2016. Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM). *Scientific Journal of Informatics*, Vol. 3(1), pp. 34-45.

[3] Oh, K., Lim, C., Kim, S., & Choi, H. 2013. Research Tren Analysis using Word Similarities and Clusters. *International Journal of Multimedia and Ubiquitous Engineering,*Vol. 9(1), pp. 185-196.

[4] Robertson, S. 2004. Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation*, Vol. 60(5), pp. 503-520.

[5] Wulandini, F., & Nugroho, A. S. 2009. Text Classification Using Support Vector Machine for Web mining Based Spation Temporal Analysis of the Spread of Tropical Diseases. In *International Conference on Rural Information and Communication Technology*, pp. 189-192.

[6] Pressman, R. S. 2002. *Rekayasa Perangkat Lunak Pendekatan Praktisi (Buku Satu)*. Yogyakarta: Andi Offset.

[7] Hetami, A. (2015). Perancangan Information Retrieval (IR) Untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris dengn Pembobotan Vector Space Model. *Jurnal Ilmiah Teknologi Informasi Asia,* Vol 9(1), pp. 53-59.

[8] Han, J., Kamber, M., & Pei, J. 2012. *Data Mining: Concepts and Techniques Third Edition.* Waltham, MA: Morgan Kaufmann, pp. 1-703.

[9] Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. 2012. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data applications.* Oxford: Elsevier, pp. 1-1025.

[10] Zhai, C., & Aggarwal, C. C. 2012. *Mining Text Data.* New York: Springer.

[11] Feldman, R., & Sanger, J. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

[12] Adi, C. K. 2016. An Identification of Tuberculosis (Tb) Disease in Humans using Naïve Bayesian Method. *Scientific Journal of Informatics*, Vol. 3(2), pp. 1-10.
[13] Wijaya, M. R., Saptono, R., & Doewes, A. 2016. The Effect of Best First and Spreadsubsample on Selection of a Feature Wrapper With Naïve Bayes Classifier for The Classification of the Ratio of Inpatients. *Scientific Journal of Informatics*, Vol. 3(2), pp. 41-50.