The English Language Scientific Literature Classification Based on Abstract Using Rocchio Algorithm

U Andayani, D Arisandi, Misbah Hasugian, M F Syahputra, B Siregar

1Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Indonesia

ulfi.andayani@usu.ac.id | dedy.arisandy@usu.ac.id

Abstract. The need for the documents in the form of journals or scientific articles is currently increasing which means the need of documents available also increasing, in another side it makes difficult to find and present journals. Therefore, a method is needed to classify journals automatically according to the categories of the journal. One method that can help organize documents according to their categories is classification. In this study, Rocchio algorithm is used as a method to classify journals. The data used comes from Google Scholar. Journal consists of 4 (four) categories: Computer Vision and Pattern Recognition, Artificial Intelligent, Data Mining and Analysis, and Computer System. System testing is carried out by taking 160 journals for training data and 48 journals for testing data. The results of this study produced a web-based journal classification system and showed the method used was able to classify journals with accuracy result is 93%.

1. Introduction
A journal is an article compiled in order to give contribution to the theory or application of science or a quote from a research in the journal there are important points from the research. In general, journals are used as reference material, and research’s literature studies. Currently, the number of journal documents or scientific articles are increasing and diverse in line with the development of the internet and technology. With ease in accessing and disseminating information, the number of available documents is increasing and the need for journal documents is increasing. If the number of documents is increasing, the process of searching and presenting documents is also more difficult to do, it will be easier if the documents are classified according to their respective categories. Therefore, a method is needed to classify journals automatically according to the categories in the journal. Categories that describe the subject matter in general and represent the topic or contents of the journal without having to read it in its entirety. Part of the theory that describe text mining explain that text mining is a new and exciting area of computer science research that tries to solve the crisis of information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval, and knowledge management [8].

One method that can help organize documents according to their categories is classification. Classify documents by grouping documents that have the same content and topics into the same category [1]. Journal classifications can help the process of presenting journals according to journal categories without having to read them in their entirety and facilitate the search for a journal. By knowing the categories of journals to be searched, the search process is only carried out on documents that have the category. Previous research has been conducted to build focused crawler by classifying web pages [2][3]. They use Naïve Bayes algorithm to classify the web pages. Another research also uses Naïve
Bayes to classify the customer care twitter posting [4]. Besides utilizing Naïve Bayes, the clustering method such as Self-Organizing map (SOM) can be used to categorize articles [5].

2. Methodology

The methodology proposed for the journal classification based on the abstract consists of several steps. These steps begin with the collection of journals’ data from four categories, which are Computer Vision and Pattern Recognition, Artificial Intelligent, Data Mining and Analysis, and Computer System from Google Scholar which will be used for data training and data test. The pre-processing step to prepare documents into data that will be processed for the classification process consisting of tokenization, stop word removal and stemming. Then the value of each term or weighting is given to the terms that exist in each TFIDF weighting journal. The next step is the classification process with Rocchio algorithm which compares the results of data training and data testing to get the classification results. After the steps are passed, the results will be obtained from the classification of the four categories of journals. The general architecture of the system to be built can be seen in Figure 1.

The general architecture of the system to be built can be seen in Figure 1.

![Figure 1. General architecture.](image)

The description of the general architecture:

2.1. Preprocessing

At this step aims to synchronize and prepare the text or journal documents into data that will be processed into data for the classification process. This step has several steps, that are: tokenization, stop word removal, and finally the stemming process [6].

- **Tokenization**
  Before words are separated from sentences, they are first cleaned of punctuation marks, numbers, and symbols. This process is carried out before the tokenization process to minimize the results of the tokenization. In the tokenization process the document will be read in the form of text and then the process of cutting the input string will be carried out based on each word that composes it.

- **Stop word removal**
  The stop-word removal process is the process of deleting terms that have no meaning or are not relevant. Stop-words removal is a process for removing irrelevant words from the results of parsing a text document by comparing it to the existing stop-word.
• Stemming
  Stemming is the process of finding and changing a word into a morpheme or a basic word that builds it. At this step, the suffix words are removed at each record, without changing the meaning of the word.

• Porter Stemming
  Porter stemming is the search for the basic word of an affixed word by removing the suffix (suffix) in the English words because English does not recognize the prefix. The Porter algorithm converts all words with the same word root into a single word form (the Stem), removing the suffix that added to the back of a word from the basic word omitted and this approaches the basic conflation form [7]. The definitions used in this algorithm are:

  V  Vocal
  C  Consonant
  M  Measure (ve)
  *S  -stem ended with ’s’
  *V*  -stem that has vocal
  *d  -stem that ended with double consonant
  *o  -stem that ended with cvc, with the second last is not W, X, Y.

2.2. TF-IDF Weighting
  Term weighting is the process of assigning value to each term that exists in each journal that has passed the pre-processing process. Giving value or weight to this term uses the TF-IDF method (term frequency-Inverse document frequency). This method will calculate Term Frequency (TF) and Inverse Document Frequency (IDF) values on each token (word) from journal entries. This method is the most commonly used method in assigning weight to a term. This weighting aims to provide a value to a term in which the value of the term will be used as input to the classification process. TF-IDF is the weight value of a word taken from the TF value and the inverse DF value, which is defined by equation 1 and 2:

\[
TF - IDF (w, d) = TF (w, d) \times IDF(W)
\]

information:
TF-IDF : the weight of a word in the entire document
W : a word
d : a document
Tf (w, d) : the frequency word w occurrence in a document

\[
IDF (w) = \log\left(\frac{N}{DF(w)}\right)
\]

information:
IDF(w) : DF’s inverse from the word ‘w’
N : total of documents
DF (w) : total of documents that has the word ‘w’

2.3. Classification using algoritma Rocchio
  Rocchio’s classification method compares the content similarity between data training and data test by presenting all data into vector. Each word’s weight is a dimension in vector space. Similarity is calculated from the proximity of the angle formed between the weight of the training data and the weight
of the test data. Where only the relevant document is selected, the query point moves closer to the center of the document that is considered relevant. To calculate the weight of each word in the document used TFIDF weighting scheme. Because the heuristic / main component of Rocchio’s classification is a TFIDF weighting scheme, Rocchio's learning method is also called TFIDF Classifier [9]. In this study the test data are journal documents. There are two steps in the classification of documents. The first step is the training of documents that are already known in the category and the second step is the process of classifying journal documents that are not yet known.

Rocchio’s classification applies clusters (Centroid) in each class. Each class’s Centroid is the average of each document vector in each class / category. A class’s Centroid is the average of all vectors in class c. Centroid values are obtained by calculating the average vector for all documents [10]. Centroid class c is calculated by the equation 3:

\[ \mu(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d) \]  

Information:
- \( \vec{v}(c) \) is every document that is in class c.
- \( \vec{v}(d) \) is the representation of vector space from d.

After calculating the result of weighting and centroid’s value, carried out classification step to get the category result towards the tested journal documents, with the equation 4:

\[ Q = \sum_{t=1}^{n_{ref}} \mathbf{w}_t \mathbf{T}_t \]  

Where \( \mathbf{w}_t \) is relevancy term from the query. The classification result (Q) is the highest value at the time of the journal test.

3. Result and discussion
In this study the system has 2 steps that are the process of data training step and data test step. Each journal that is successfully inputted will go through these two steps. The first step is data training that is the classification step or to build a model or function for a journal whose labels or categories are already known. The data training phase can be seen in Figure 2.

The next step is carried out system testing to determine the accuracy of the model or function that was built during the data training phase. The data test phase is to classify journals that are not yet known in their category. The journals that used in the test data phase are 48 with 12 journals in each category. System testing is done. Data steps can be seen and figure 3.

System testing is carried out to determine the ability of the system that built in classifying 4 types of journal categories.
Examples of training data can be seen in Figure 4 and after processing the data testing, the journals that have not been categorized will be categorized according to their categories. Journal testing results can be seen in Figure 5.
4. Conclusion and future work
After conducting literature studies, design, analysis, implementation and testing of an abstract based journal classification system can be summarized as follows:

- Rocchio algorithm can be used for classification of journals with 93% accuracy.
- The more the number of words stored, the more the word’s weight will be compared to the text test. However, the longer it will take time for the document classification process.
- The error that happen on document classification is because there is the same word and the weighting value of the word from another category is greater.

For the next research, it is expected that this system is not only classifying English-language journals in 4 (four) categories but can also be used to classify documents in all fields and other documents such as online news.

References
[1] Manning C, Raghavan P and Schütze H 2009 An Introduction to Information Retrieval (Cambridge: Cambridge University Press)
[2] Amalia A, Gunawan D, Najwan A and Meirina F 2016 Focused crawler for the acquisition of health articles 2016 International Conference on Data and Software Engineering (ICoDSE) pp 1–6
[3] Gunawan D, Amalia A and Najwan A 2017 Improving Data Collection on Article Clustering by Using Distributed Focused Crawler J. Computing and Applied Informatics (JoCAI) 1 pp 1–12
[4] D Gunawan et al 2018 J. Phys.: Conf. Ser. 978 012119
[5] Gunawan D, Amalia A and Charisma I 2017 Clustering articles in bahasa indonesia using self-organizing map 2017 International Conference on Electrical Engineering and Informatics (ICELTICs) pp 239–244
[6] Korde V 2012 Text Classification and Classifiers A Survey International Journal Of Artificial Intelligence and Applications (Ijaia) pp 85-99
[7] Peter W 2006 The Porter Stemming Algorithm: Then and Now Program: Electronic Library And Information Systems pp 219-223
[8] Feldman R and Sanger J 2007 The Text Mining Handbook Advanced Approaches In Analyzing Unstructured Data (New York : Cambridge University Press)
[9] Joachims T 1997 A Probabilistic Analysis of The Rocchio Algorithm With Tfidf For Text Categorization Proceeding Icml Proceedings Of The Fourteenth International Conference On Machine Learning pp 143-151
[10] Gao G and Guan S 2012 Text Categorization Based on Improved Rocchio Algorithm International Conference On Systems And Informatics (Icsai) 2012 pp 2247-2250