Scientific Documents Classification Using Support Vector Machine Algorithm

To cite this article: I Jaya et al 2019 J. Phys.: Conf. Ser. 1235 012082

View the article online for updates and enhancements.
Scientific Documents Classification Using Support Vector Machine Algorithm

I Jaya, I Aulia, S M Hardi, J T Tarigan, M S Lydia and Caroline
Faculty of Computer Science and Information Technology, Universitas Sumatera Utara
Email: ivanjaya@usu.ac.id

Abstract. One of the ways to publish scientific paper is to present it in a conference. Before being presented in the conference, the registered scientific paper has to be grouped into categories. Because there are so many scientific papers that have been registered, it is an urgent need to establish a system that can automatically identify the topic of scientific papers and classify them according to recognized topics. One method that can be used to organize documents according to their categories is classification. In this study, Support Vector Machine algorithm is used as a method to classify documents into 5 categories of computer science. System test is done by taking 150 documents as data sample and 50 documents as data testing. This study produced a classification system of scientific documents and showed the method used was able to classify documents with 90% accuracy results.

1. Introduction
Research needs to be disseminated to provide updates in the field of science and exchange information with the research community. By publishing the research, the researcher get feedback from experts/specialists at the research field. Publication also performed in order to get copyright for the researcher. Furthermore, the result of research also disseminated so that people can use it for research interest, both as reference, research development or implementation in real. One way to publish scientific paper is following the conference.

Conference is a formal meeting between group or academic organizations to discuss together the latest information or problems at the research field [1]. Before being presented in the conference, the registered scientific paper was grouped into categories. The conference organizer usually grouping scientific papers manually by reading the summary (abstract) each of them. The more scientific papers were accepted the longer process of grouping them. Therefore we need the system that can identify the scientific papers’s topic and group them according to the topic that can be recognized automatically. The grouping is performed by classification method.

In this research, document classification will be performed by recognizing documents’ category through the documents’ abstract. The document’s abstract part will go through several preprocessing phases. Afterwards the result of the process will enter to word weighting phase and then go into classification phase using Support Vector Machine algorithm.

2. Previous Research
The previous document classification research had been done using Naive Bayes algorithm which is added with Neural Network approachment. The result is better than the classification without neural network [2].
Rusdi Efendi uses Naive Bayes algorithm to classify report documents with 30 training data of documents and 30 testing data of documents. The research produces 86.67% accuracy [3]. Another research uses Naïve Bayes to divide twitter input into several groups [4]. Meanwhile, other researches utilize Naive Bayes to build focused crawler [5][6].

Furthermore, document classification had been done using TF-IDF that has been modified on weighting phase. In this research the same term will have different weight depends on where the term is. Then the documents are classified using Naive Bayes algorithm and Frequency Ratio Accumulation Method [7]. In addition, TF-IDF is also used for document clustering [8].

Next research using K-Nearest Neighbour to classify based on winnowing fingerprint. Winnowing is an algorithm that is used to detect the similarity of text document’s contents by breaking sentences in it into some characters as long as k-grams and producing output which are collection of hash values called fingerprint. In this research fingerprint is used as characteristic of text document then group the text documents based on that characteristic [9].

Misbah Hasugian had build the journal classification system based on abstract using Rocchio algorithm. The method compares the similarity of training data contents and testing data contents. The method is able to classify journals with 93% accuracy [10].

The research with Support Vector Machine is used by Adyatma Bhaskara to classify report documents. That research was successful in classifying report documents using 3110 x 1716 training data and 1252 x 1716 testing data [11].

3. Method

The method that was proposed to classify the scientific papers consists of several processes. The processes that will be performed are illustrated in Figure 2 as follows: The scientific document (data set) is inputted based on metadata and content. Metadata consists of title, researcher’s name and keywords and for the process abstract is inputted based on its contents. After the input process, data will go through pre-processing text and weighting process using TF-IDF method. Classification process with Support Vector Machine algorithm is used in this phase to classify scientific papers into categories based on the topic of each document. Both training and testing do the same process. The different between training and testing is at the training the result become a model where the result values will be compare to the data testing after processing with SVM.

A. Data Set

Data that it used for this research consists of training data and testing data. Data is scientific documents with .pdf document format taken from International Conference on Computing and Applied Informatics (ICCAI) and Springerlink collection. The data will be processed to produce document classification in several categories that are:

- Computer Graphics & Image Vision
- Computer System
- Data Mining
- Human Computer Interaction
- Information Security

The part of document that will be processed is abstract part. The number of training data for each category is 30 where the average words in the abstract is 150 to 200.

B. Text Preprocessing

Input in the form of abstract text is data that haven’t structured yet. To get structured data text required preprocessing phase. Preprocessing is performed to remove parts or text that’s not needed in order to get qualified data to be processed [11]. There are several phases that is performed on preprocessing that are:

1. Tokenization

   Tokenization is a phase of changing the string into term that configure it. On tokenization process, space and punctuation is used to perform token fission for each document. Words are seperated from the original text without seeing duplication. In this process term which typed integer and punctuation will be removed.
2. Stopword Removal
Stopword removal is term or words disposal that are considered meaningless or irrelevant. Stopword must be deleted from a text because it can make text harder and less important to text mining process. Stopword in English is like conjunction, preposition or adverb.

3. Stemming
On stemming phase the word is changed into basic word or the process to search root word from each word that’s delivered from stopword removal. Searching the basic form of a word is performed by deleting affix in the word. This research using Porter algorithm to stemming process. Porter algorithm consists of five word reduction phases, applied sequentially. For each phase there are many convention to choose rules such as rules selection from each rule group that are applied to the longest suffix.

C. TFIDF Weighting
Term Frequency–Inverse Document Frequency (TFIDF) is the method that is used to measure how important a word of a document. TFIDF value increases proportionately by how many times a word appears in document, but contrary to word frequency in the corpus. TF (Term Frequency) asserts the number of documents that contains a word in one publication segment. TFIDF is the weight value of a word taken from TF values and DF inverse values [12], defined by:

$$\text{IDF} (w) = \log \left( \frac{N}{\text{DF}(w)} \right)$$ (1)

$$\text{TFIDF}(w,d) = \text{TF}(w,d) \times \text{IDF}(w)$$ (2)

TF-IDF (w,d) : weight of the word in entire document
w : the word
d : the document
TF(w,d) : frequency of w word appearance in document
IDF(w) : inverse DF from w word
N : the number of all documents
DF (w) : the number of document that contains w word

If the value of N = DF(w), then it will get 0 (zero) result for IDF calculation. For that, it can be added 1 value to IDF side.

D. Classification Support Vector Machine
Support Vector Machine (SVM) is one of technique to perform prediction, both in classification or regression. This technique include in supervised classification method because it has a certain target learning. SVM is the classification method that works by searching hyperplane with the biggest margin. Hyperplane is dividing line (decision boundary) data between classes, while margin is distance between hyperplane with the closest data for each class. The closest data to hyperplane for each class called support vector [13].

On document classification, training data from the previous process is required to make SVM model. SVM model is the classification model based on vector space. All of document vectors is mapped then search for class separator function. SVM is classification method that divide vector space into 2 parts that are positive class and negative class by hyperplane.

SVM when first introduced just be able to classify data into two classes. Further research develops SVM so it can classify data that have more that two classes has performed continuously. There is an option to implement SVM multiclass by integrating several binary SVMs or integrating all data consisting of some classes into a form of optimization problem. The second approachment in optimization problem that must be solved is more complicated. There are two method that is used
generally to implement SVM multiclass by approaching, that are “one-against-all” method and “one-against-one” method. In this research is used “one-against-one” method.

One-against-one method built k(k-1)/2 binary classification model with k is the number of classes. Each classification model is trained in data from two classes. In this research is performed classification for 5 classes so the number of binary classification model that is done is:

\[
\frac{5(5 - 1)}{2} = 10
\]

With all of binary classification which is performed is:

![Figure 1. One-against-one Method](image)

There are several methods to perform testing after all classification models built completely. One of them is voting method [14]. If x data is entered into training result function and the result asserts x is i class, so the vote for i class is plus one. Class of x data will be determined from the most votes. If there are two classes which the number of votes is same, then the class which the index is smaller will assert as class of data.

![Figure 2. General Architecture to Classify Scientific Papers](image)
4. Results and Discussions

On this phase is performed testing to the system to know if the system has worked properly or not, to measure how big the accuracy of the system in performing its function. Using SVM model from 150 training documents that divide into 5 categories, will be performed classification testing. The number of documents that is used to test are 10 documents for each category. The result of the testing, delivered as shown in the Table 1.

Table 1. The Testing Result of Documents Classification

| Category                              | The Number of Documents | True Classification |
|---------------------------------------|-------------------------|---------------------|
| Computer Graphics and Image Vision    | 10                      | 10                  |
| Computer System                       | 10                      | 9                   |
| Data Science                          | 10                      | 9                   |
| Human Computer Interaction            | 10                      | 8                   |
| Information Security                  | 10                      | 9                   |

From Table 1, it can be seen that the results of the classification are not entirely accurate, there are several documents that are different categories with the categories that should be. This is due to lack of training data which causes a lack of important words that characterize a document and there are also dominant words used in several categories at once so that errors occur in the classification of documents. The level of accuracy of classification success can be seen by calculation:

\[
\text{accuracy} = \frac{\text{the number of true classification}}{\text{the number of document classification}} \times 100\%
\]

\[
\text{accuracy} = \frac{45}{50} \times 100\%
\]

\[
= 90\%
\]

5. Conclusions

Support Vector Machine algorithm can be used to classify scientific papers based on abstracts into 5 categories: Computer Graphics and Image Vision, Computer Systems, Data Science, Human Computer Interaction and Information Security with 90% accuracy. The error when document classification inappropriate is because of the same words between the document and the model, where values of the weighted words from other categories are bigger. Because of that, classification of the inputted document is closer to other categories.

6. Acknowledgments

The authors gratefully acknowledge that the present research is supported by Fund Dissemination IPTEKS Research Results for Lecturers / Researchers Universitas Sumatera Utara.

References
[1] Jatmiko W, Wibisono A, Wisesa H A and A’unina Q 2016 Prosedur Penyelenggaraan Konferensi Internasional (Jakarta: Universitas Indonesia).
[2] Zhang M, Gao X, Cao M D and Ma Y 2006 Neural Networks for Scientific Paper Classification Proc. of 1st Int. Conf on Innovative Computing, Information and Control (ICICIC’06).
[3] Efendi R., Firsandaya R M, Mila S J 2012 Klasifikasi Dokumen Berbahasa Indonesia Menggunakan Naive Bayes Classifier J. of Research in Comp. Science and Applications 1 1.
[4] D Gunawan et al 2018 J. Phys.: Conf. Ser. 978 012119
[5] 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
[6] 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
[7] D‘cunha A, Sen A K 2015 Hierarchical Approach for Scientific Document Classification. Int. Conf. on Computing, Communication and Automation (ICCCA2015) pp 100-103.
[8] 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
[9] Sanjaya S, Absar E A 2015 Pengelompokan Dokumen Menggunakan Winnowing Fingerprint dengan Metode K-Nearest Neighbour Jurnal CoreIT 1 2.
[10] Hasugian M 2017 Klasifikasi Jurnal Berbahasa Inggris Berdasarkan Abstrak dengan Algoritma Rocchio (Medan : Universitas Sumatera Utara).
[11] Hutomo A B 2014 Klasifikasi Dokumen Berita Menggunakan Metode Support Vector Machine dengan Kernel Radial Basis Function (Bogor:Institut Pertanian Bogor).
[12] Feldman R, Sanger J 2007 The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data (New York : Cambridge University Press).
[13] Nugroho A S 2008 Support Vector Machine: Paradigma Baru dalam Softcomputing. Konferensi Nasional Sistem dan Informatika pp 92-99.
[14] Chih-Wei H and Chih-Jen L 2002 A Comparison of Methods for Multi-class Support Vector Machines. Department of Computer Science and Information Engineering (Taiwan:National Taiwan University).