Advanced Science and Technology Letters
Vol.111 (COMCOMS 2015), pp.263-266
http://dx.doi.org/10.14257/astl.2015.111.50

Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier

Jong-Yeol Yoo[1] and Dongmin Yang[1]

[1]Dept. Of Information & Communications Engineering,
Daejeon University, Daejeon, Korea
gum10011@naver.com, dmyang@dju.kr

Abstract. Recently due to large-scale data spread in digital economy, the era of big data is coming. Through big data, unstructured text data consisting of technical text document, confidential document, false information documents are experiencing serious problems in the runoff. To prevent this, the need of art to sort and process the document consisting of text data has increased. In this paper, we propose a novel text classification scheme which learns some data sets and correctly classifies unstructured text data into two different categories, True and False. The proposed method is implemented using Naïve Bayes document classifier and TF-IDF.

Keywords: Document Classification, Naïve Bayes, TF-IDF

1 Introduction

With the development of IT technology, people were exposed to numerous data all the time. However, these data may be adversely effective to the individual or used as a security threat by malicious users. For winnowing meaningful data or excluding data including security risk, content filtering is a key technology in the Big Data era [1-5]. In addition, the existing content filtering algorithms used in structured data processing is not available in dealing with unstructured text data collected at a tremendous rate. Therefore, the filtering performance must be improved to extract meaningful data and filter the content from a huge data in real time. It is very useful to companies, which is interesting in consumers' preferences and countries which must provide reliable security services to people. In this paper, we propose a novel text classification scheme which learns some data sets and correctly classifies text data into two different categories, True and False. The training phase of the text classification consists of two steps. At the first step, a document is entered and the morphological analysis has been performed. In the next training step, TF-IDF features have been extracted and conditional probabilities for each class of two categories, True and False, have been learned by Naive Bayes classifier. The trained classifier determines whether an input document is True or False. To verify its practicality, we implement the text classifier using python libraries.

This paper is organized as follows. Section II reviews the related works. Section III presents a text classification scheme. Finally, the conclusions are presented in Section

ISSN: 2287-1233 ASTL

IV.

2 Related Works

2.1 TF-IDF (Term Frequency-Inverse Document Frequency) [6]

TF-IDF is a weighting factor used in information and text mining. It represents a numerical statistic which rejects how important a word is to a document in a collection or corpus. TF (Term Frequency) is a value that represents how often a particular word appears in a document. The more frequent the value of the word is, the more important it is in the document. However, if the word appears too frequently in the corpus, this means that the word is so common in the corpus. For example, \atom" could be the key word because it does not appear in normal documents. But it becomes a common word in physics in which it is mentioned frequently. This situation can be adjusted by IDF (Inverse Document Frequency). DF (Document Frequency) can be obtained by dividing the total number of documents by the number of documents containing the term. IDF is the inverse value of DF. TF-IDF is the product of two statistics, TF and IDF. A high weight in TF-IDF is reached by a high term frequency and a low document frequency of the term in the whole collection of documents. Hence, the weights tend to filter out common terms. TF-IDF can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

2.2 Naïve Bayes Classifier [7-11]

Naïve Bayes classifier is a kind of probabilistic classifier based on applying Bayes' theorem. It assumes that all features are strong independent. Naïve Bayes classifier has the following three advantages. First, in some probability models, Naïve Bayes classifier can be effectively trained in supervised learning environment. Second, the amount of training data used to estimate the necessary parameters for classification is small. Third, despite a simple design, Naïve Bayes classifier operates well in various and complicated situations. Let's suppose that the parameters of a document are x, y and z, and all documents are divided into two categories, A and B. And $P_A(x, y, z)$ is the probability that a document belongs to A. If $P_A(x, y, z) > P_B(x, y, z)$, the document is in A, otherwise, it is in B.

3 Text Classification using TF-IDF and Naïve Bayes Classifier

The text classifier is very simple, which consists of two phase: training and test phases as shown in Fig. 1. The number of document categories are two, True and

False. Two categories are in general. For example, spam or not, confidential or not, and etc.

As the first step of training phase, morphological analysis is performed. It is the identification, analysis and description of the structure of a given language's morphemes and other linguistic units. We implement it using python library functions, PyStemmer. After the morphological analysis, TF-IDF values are computed using python library functions, tfidf, from github. At the final step of training phase, the text classifier is trained by two sets of data, True and False. In this paper, two sets of spam and ham email corpus, Enron-spam, is used, which could be obtained from [12].

In the testing phase, if a document is entered, morphological analysis, feature extraction by TF-IDF and classification by Naïve Bayes Classifier are performed. When $P_{Spam}(TF\text{-}IDF_{DOC}) > P_{Ham}(TF\text{-}IDF_{DOC})$, the document is in *Spam*, otherwise, it is in *Ham*.

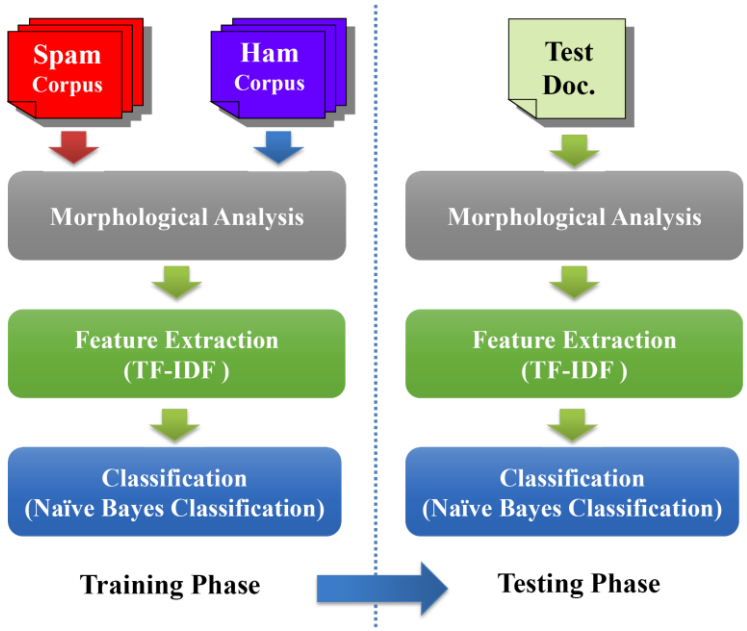

Fig. 1. Procedure of Unstructured Text Classification

4 Conclusion

In this paper, we propose a novel text classification scheme which learns a set of documents and correctly classifies text data into two categories. In addition, we implement the text classifier using python libraries. For future works, instead of Naïve Bayes classifier, some machine learning techniques such as SVM(Support Vector Machine) and word embedding a will be applied for more accuracy.

References

1. Ismaila Idris: E-mail Spam Classification with Artificial Neural Network and Negative Selection Algorithm. International Journal of Computer Science & Communication Networks, 1(3), 227-231 (2014)
2. Alaa El-Halees: Filtering Spam E-Mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques. The International Arab Journal of Information Technology, 6(1), 52-59 (2009)
3. Seongwook Youn, Dennis McLeod: A Comparative Study for Email Classification. In: Advances and Innovations in Systems, Computing Sciences and Software Engineering, LNCS, vol. pp. 387-391. Springer, Heidelberg (2007)
4. Istvan Pilaszy: Text Categorization and Support Vector Machines. In: the 6th International Symposium of Hungarian Researchers on Computational Intelligence. (2005)
5. Nishtha Jatana, Kapil Sharma: Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach. In: 2014 International Conference on Computing for Sustainable Global Development (INDIACom), pp. 939-942. IEEE Press, New York (2014)
6. TF-IDF, https://en.wikipedia.org/wiki/Tf-idf
7. Naive Bayes classifier, https://en.wikipedia.org/wiki/Naive_Bayes_classifier
8. LI Aiwu, LIU Hongying: Utilizing Improved Bayesian Algorithm to Identify Blog Comment Spam. In: 2012 IEEE Symposium on Robotics and Applications(ISRA), pp. 423-426 (2012)
9. Dae-Ha Park, Eun-Ae Cho, Byung-Won On: Social Spam Discovery using Bayesian Network Classifiers based on Feature Extractions. In: 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp. 1808-1811. IEEE Press, New York (2013)
10. Junyan Peng,, Patrick P. K. Chan: Revised Naive Bayes CL Assifier for Combating the Focus Attack in Spam Filtering. In: the 2013 International Conference on Machine Learning and Cybernetics, pp. 610-614. (2013)
11. M. Tariq Banday, Shafiya Afzal Sheikh: Multilingual E-mail Classification using Bayesian Filtering and Language Translation. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I), pp. 696-701. IEEE Press, New York (2014)
12. Enron-Spam, http://www.aueb.gr/users/ion/data/enron-spam/
13. Python, https://www.python.org

