International Journal of
Applied Mathematics,
Electronics and Computers

ISSN: 2147-8228 http://ijamec. atscience.org

Advanced Technology and Science

Original Research Paper

Classification Performance of the Different Stemming Methods

Mehmet Balcı [1], Rıdvan Saraçoğlu [2], Şakir Taşdemir *[1], Adem Gölcük [3]

Accepted 15th August 2014 **DOI: 10.18100/ijamec.27805**

Abstract: Saving textual data and accessing them in many fields have become one of the basic problems nowadays. The usage of these data effectively is directly related to the development of storage and access tools that will be used. Therefore, software programs using different methods have been developed. One of the points that need to be taken into account is data classifying. Because using raw data in these classifying processes is harmful, finding the stem of the texts is useful. In this study, the successes of two different stemming algorithms in the text classifying are comparatively examined.

Keywords: *Text Processing, Classification, kNN, Stemming*

1. Introduction

Technology that is a part of our daily lives brings an outburst of data. The most important examples for the increasing data at this speed are electronic mails and web sites [1]. This available raw data should be saved, processed through appropriate methods and they should be reached to the users. Data mining coming out because of these necessities can be described as retrieval of the useful information from the great amount of data [2].

Text mining is method of the data mining. Text mining is commonly used in order to find the documents written in the same subject, the ones related to each other and discover the relations among concepts [3].

Text processing techniques are used in finding the texts written especially in the same subject or disciplinary and classifying the newly-come data according to the available fields. In order to make a more successful classifying, stemming existing in the preprocessing of text mining carries great importance. The process of stemming differs according to the languages. For instance, it is possible to develop a stemmer for a language that has less affixes like English by just looking the glossary of affixes up [4]. Because Turkish is an agglutinate language, the number of affixes and the varieties of addition make it necessary to examine in a detailed way [5].

1.1. Text Classification

Text classifying is the name given to the process of assigning the written documents to the particular classes according to their contents. The process of separating the news coming from a source according to their subjects can be given as an example for the text classifying [6]. There are different methods used in the text classifying. The first of the methods that can be applied for the text classifying is the approach of knowledge engineering [7]. In this method, classifying rules are formed by specialists and newly-coming documents can be classified according to these rules. That the classifying rules are formed by specialists by hand will be a difficult and time-consuming process. In order to form the inquiries with the aim of classifying by editors, two-day period may be needed [8]. For this reason, this method will be very ineffective and inadequate for many application fields. For instance, determining the rules for the classes in a situation where there are many classes can be difficult. What is more, specialist knowledge will not exist for a user who wants to classify their documents. Furthermore, revising and reforming rules will be needed in the situation of change of classes [6].

In this study, the method of k-Nearest Neighbors algorithm that is one of the machine learning was used. For this, the mathematical model of texts that will be classified beforehand.

When examined this process closely, the obligation of using the documents with the state of marker set reflecting the contents of the documents not with their states that are presented to the system is noticed. These markers of content are given names such as keyword, index word and definer. Using markers of content was firstly suggested by Luhn at the end of the 1950s [9].

2. Material and Method

2.1. The Mathematical Model of the Texts-Vector Space Model

Modeling the texts mathematically needs to be done in order for textual data to be processed by computer systems and to apply data mining on the texts. The vector space model is a model considering the documents as a term vector by weighting the terms consisting of the documents [10].

The vector space model symbolizes a special meaning of natural language documents in the multi-dimensional space. In the modelling of the texts, the most common used model is the vector space model.

[1] *Computer Technologies Department, Higher School of Vocational and Technical Sciences, Selçuk University, Konya/Turkey*
[2] *Electrical and Electronic Engineering Department, Faculty of Engineering and Architecture, Yüzüncü Yıl University, Van/Turkey*
[3] *Computer Technologies Department, Higher School of Vocational and Technical Sciences, Karamanoglu Mehmetbey University, Karaman/Turkey*
** Corresponding Author: Email: stasdemir@selcuk.edu.tr*
 # This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

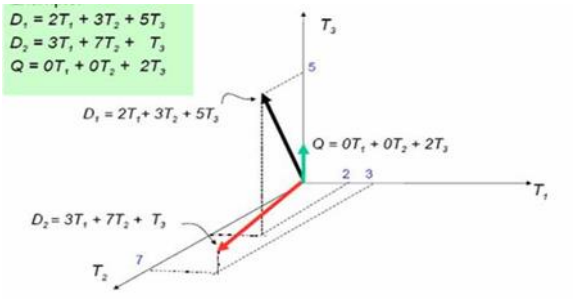

Figure 1. Vector space model

In the vector space model, each vector is represented by a vector. While a vector of a document is being formed, all the terms are stemmed first and the following steps are being followed respectively:

- The terms that the documents include and their frequencies are determined.
- A nonrecurring dictionary is formed from the terms in the documents.
- Each document is brought to a vector form according to the terms they contain.
- The vectors of all the documents are kept as an array.

2.2. The Similarity Between the Documents

In order to find the similarities of two documents to each other, the mathematical model (vectors) of the documents need to be done first. Then, the distances of these vectors to each other are calculated. The simplest approach is to find Euclidean distance between two documents related to documents [2]. This method is used in the similarity-based search and document classifying. The Euclidean distance between two vectors of documents is found through the formula below:

$$Euclidean(d_q, d_i) = \sqrt{\left[\sum_{j=1}^{M} (w_{q,j} - w_{i,j})^2\right]} \qquad (1)$$

By calculating inquiry vectors searched for the similarity-based searches and Euclidean distances of vectors related to the documents in the document collection, the document having least distance to the inquiry is retrieved as the document resembling most.

As for the document classifying, a range of processing determining whether the documents are in the same class or not is made by finding Euclidean distances of all the document vectors in the document collection.

2.3. k-NN(Nearest Neighbors) Algorithm

Classification of objects is an important area of research and application in a variety of fields [11]. This algorithm is the controlled learning algorithm that is a result of classifying the inquiry vector with the vector in the k-Nearest Neighbors. In order to classify this algorithm with a new vector, the vectors of the document and education documents. An inquiry sample is given and the education point as pieces of K nearest to this inquiry point is found. As for classifying, it is made with the one that is most of these objects as pieces of k. k-NN application is a neighborliness classifying algorithm used to classify new inquiry sample [12].

Nearest Neighbor based classifiers saves all the documents in the education group during the education in their memories. When a document has come for classifying, the classifier chooses the Nearest Neighbor pieces of k to this document. Then, by looking at the classes of Neighbor documents, they are assigned to one or more classes this document [6].

In order for k-NN algorithm to work, it is essential that a relation of similarity between documents exists. This relation of similarity is found by calculating the Euclidean distances between vectors or Cosine similarities.

2.4. Longest Match Algorithm

In this method, firstly the word is searched in a dictionary where the stems of the words related to the document exist. If it cannot be found, the process of searching is made again by deleting a letter from the end of the word. The process ends when a stem is found or a word is a letter. Although this method is one of the most understandable methods in terms of application, it has a bad performance in terms of time because many times of searching in the dictionary are made for each term during stemming [13].

2.5. Fixed Length Algorithm

While finding the stem of each term that will be stemming in this method, letters of fixed amount are considered as stems. The researchers using this method make researches by accepting the different amounts of letters as stems. In the experimental research that will be mentioned below, this stemming method is tried in four different ways by stemming with letters of 4, 5, 6 and 7 and their performances on classification are examined [13].

2.6. Dataset

In the data set used in the study, there are 1000 documents including many dissertations and articles. Because that all the text contents belonging to these documents exist in the data set causes an awkward structure in the processing the processes in software and it causes more waste of time, marker sets presenting the contents of the documents best are used in the data set. As marker set, the name of document, abstract and keywords are chosen. After the documents in the data set undergo preprocessing processes such as, decomposition (removing the punctuation marks, converting all the letters into lower cases.), removing the stop word. They are separated word by word and prepared for the stemming. Some numeric data belonging to the data set are seen in Table 1.

Table 1. Numeric data of dataset

Data Set Information	Count
Number of Document	1000
Total Number of Words	327.636
Number of Stop Words	61.454
Total Raw Term Count (After discarding stop words)	266.182

3. Practice

In this study, the effect of stemming upon the classifying is also researched with the aim of comparing the stemming methods. In this matter, k-NN algorithm from the text classifying algorithms is used. Because this algorithm is based on the principle of closeness of the documents to each other, in the documents whose Euclidean distances are calculated and that are mentioned above "n Fold Cross Validation" method is used as test and educational documents.

3.1. n Fold Cross Validation

Data set is randomly divided as n group. While 1st group is separated for the test, the model is set with the other groups. The

set model is used for the data separated for the model test. The process repeats n times and respectively all the groups (the whole of the data set) have been used for the test.

3.2. The Classifying Successes of Stemming Methods

The Vector Space and Forming the Models of Mathematical Texts, weighting Processes, Calculation of Euclidean Distances, The Stages of Application of k-NN Algorithm which are all mentioned above are separately applied for the applications of both LMA and FLA's usages of 4, 5, 6 and 7 letters. In our study, 607 documents related to 5 classes (Educational Sciences, Physics, Vegetable and Fruit Food, The Management, Economics) mentioned most from the 1000 documents related to 15 classes with the classifier are chosen. After the closest document in terms of Euclidean distance of each document vector is determined, test document is assigned to the class of the educational document that is closest to itself. The education process ends when all the documents are completed to be classified with k-NN.

Over the documents chosen, cross validation is applied by choosing n=8 and as a result of cross validation, the successes of classifying occur as seen in the Table 2 below:

Table 2. The Success Rations of k-NN Classifying of the Stemming

Stemming Method	k parameter value	Performance (%)
Longest Match	9	72
Fixed Length (4 let.)	8	64
Fixed Length (5 let.)	5	60
Fixed Length (6 let.)	9	68
Fixed Length (7 let.)	9	72

4. Conclusion

As a result of k-NN classifying application, the one giving the most successful result is Longest Match Algorithm. From the lengths used in Fixed Length Algorithm, the most successful classifying is seven letter, having the same classifying success with Longest Match. In the same classifying process, the one having the least classifying success is found to be five-letter stemming method. As understood from the Table 2 above, while the increase of length for five, six and seven letters increases the classifying success, the same situation does not occur for the 4 letters. The classifying success and the stem length used in the Fixed Length stemming is not directly proportional. The reason for the fact that classifying success is not completely directly proportional with stem length is related to the additions the raw terms in the data set get (derivational and inflectional affixes) and occurred grammatical events as well as determining of the class information to which the documents belong in the document collection manually without any specialist views.

As a result, by looking at the classifying success rations of k-NN, this can be said "Stemming of seven letters of Fixed Length Algorithm gives the closest result to Longest Match Algorithm." [13]

Acknowledgements

This study was taken from the dissertation named as "Comparative Analysis of the Longest Match Algorithm in Computer Based Text Processing" and prepared in The Graduate School of Natural And Applied Science of Selcuk University in 2010.

References

[1] Kantardzic M., 2003, Data Mining:Concepts, Models, Methods, and Algorithms, IEEE Pres, Wiley Interscience Publications.

[2] Saracoğlu R., 2007, Searching For Similar Documents Using Fuzzy Clustering, PhD Thesis, Graduate School of Natural and Applied Sciences, Selçuk University, Konya.

[3] Yıldırım P.(*), Uludağ M.(**), Görür A.(*), 2008, Hastane Bilgi Sistemlerinde Veri Madenciliği, Akademik Bilişim Konferansları'08, (*) Çankaya Üniversitesi, Bilgisayar Mühendisliği Bölümü, Ankara. (**) European Bioinformatics Institute, Cambridge, UK.

[4] Porter, M.F., 1980, An Algorithm For Suffix Stripping, Program, 14(3):130-137.

[5] Jurafsky, D. and Martin, J., 2000, Speech and Language Processing, Prentice Hall, New Jersey.

[6] Kesgin F., 2007, Topıc Detectıon System For Turkısh Texts, Master Thesis, Graduate School of Natural and Applied Sciences, Istanbul Technical University,Istanbul.

[7] Joachims, T., 2002, Learning to classify text using support vector machines, Kluwer Academic Publishers, Boston.

[8] Jackson, P., Moulinier, I., 2002, Natural language processing for online applications: text retrieval, extraction, and categorization, Amsterdam.

[9] Lassila O., 1998, Web Metadata : A Matter of Semantics. IEEE Iternet Computing, pp. 30-37.

[10] Eroğlu M., 2000, A Study On The Effects Of Stemming And Thesaurus For Retrieving Information In Turkish Documents, Master Thesis, Hacettepe University, Computer Enginering, Ankara, Türkiye.

[11] Keller J.M., Gray M.R., Givens J.A., A Fuzzy K-Nearest Neighbor Algorithm, Systems, Man and Cybernetics, IEEE Transactions on Volume:SMC-15, Issue:41Page(s):580-585,1985.

[12] Pilavcılar İ., 2007, Metin Madenciliği ile Metin Sınıflandırma(KNN Algoritması) – 3, Yazılım Mühendisliği İleri Seviye Makaleleri http://www.csharpnedir.com.

[13] Balcı M., 2010 Comparative Analysis of The Longest Match Algorithm in Computer Based Text Processing, Master Thesis, Graduate School of Natural and Applied Sciences, Selçuk University, Konya.