Detection of The News about Turkey with Web-based Text Mining System

Ferruh YİĞİT
Selcuk University
Ilgın Vocational School
Konya, TURKEY
ferruhyigit@selcuk.edu.tr

Ömer Kaan BAYKAN
Selcuk University
Faculty of Engineering, Department of Computer Engineering
Konya, TURKEY
obaykan@selcuk.edu.tr

Abstract:

Developments in information and communication technology give possibility to publish the news which happens in any region of the world to long distances. Increasing in web-based news texts makes it difficult to monitor and follow the information related to subject to be interested in the news. For this reason, web based text mining is often used for classification and monitoring of web-based news text. In this study, a classification algorithm is suggested which takes the news related to Turkey from international news channel’s RSS systems (Real Simple Syndication). The algorithm is based on the term weighting and KNN algorithm. 175 texts related to Turkey, and 150 texts related to other countries have been obtained from different news channels’ RSS systems manually (Aljazeera English, China.org English, CNN.com World, Irna, The Moscow Times News, Reuters World News, Spiegel Online International, UKWorldNews) in different parts of the world. A glossary and a database is created by using the roots of the words at texts with Porter Stemmer algorithm and some other words such as TBMM (Turkish Grand National Assembly), Çankaya and so on are added to glossary. Database is used for classification in following stages. For testing purpose, 100 different texts obtained from news channels randomly. The words in the news texts determined and identified whether they are existed in the glossary or not, and a feature vector is created. K-NN algorithm is used to determine the feature vector belongs to which class. The proposed system carries out the process of classification of the news with 90% accuracy.

Keywords- Web based text mining; K-NN; Turkish News

INTRODUCTION

As a result of fast developments at World Wide Web, news started to be flown to the users by web. The rapid increase in web-based new texts make difficult to be seen, analyzed and classified by the users. Web mining techniques can be used for automatic identify and analysis of useful information from WWW. Web content mining (information discovery on www) and web usage mining (describes the research in analyzing the user access patterns from web servers) are the two headings analyzed under the web mining [2]. The processes such as language detection of the pages, determining the category and the keywords of the pages and so on are realized in web content mining.
Classification of web pages is based on detection whether it is belonging to one of predefined categories. The text in page, the content of hyperlink or both of them can be used together on the purpose of the classification of web pages [4]. Descriptive features are extracted from text and hyperlink contents in page. Web pages are classified using features by the help of techniques such as support vector machines, artificial neural networks, fuzzy logic, genetic algorithms, k-nearest neighborhood, C4.5, decision trees and so on [5, 6, 7, 8, 9, 10]. In this study, a new method is proposed in order to determine the news about Turkey which is on the international news channel automatically. Features extracted from the texts that were taken from RSS systems of the news channels are classified with the K-NN algorithm. The results show that the proposed method can be used for classifying web pages.

The rest of the paper is organized as follows. In the second part of the article, research methodology and the dataset are clarified. The results are presented in the third section and discussion is in the fourth section.

**Research Methodologies**

The parts of proposed method which is related detection of the news about Turkey are shown in Fig 1. These parts are explained in the following subsection:

![Figure 1. The parts of proposed method](image)

**Dataset**

The texts in this study are taken from RSS systems of the international broadcast news such as (Aljazeera English, China.org English, CNN.com World, Ima, The Moscow Times News, Reuters World News, Spiegel Online International, UKWorldNews) Used texts are selected and labeled as 175 of the texts regarding Turkey, and 150 about other countries. Selected texts are used in order to feature extraction and create dictionary. 100 texts are selected randomly from the news channel on the purpose of test performance of the proposed methods.

**Pre-processing**

The text taken from RSS channels or the web pages are also included the information belonged to HTML and special characters as well as relevant text. In addition, Pronouns and conjunctions which have effects decreasing performance such as "a", "able", "about" are taken place. Such information needs to be cleaned in the texts. Firstly, HTML tags, script tags, special characters, punctuation marks and numbers found in the texts are removed. Secondly, stop words in
texts (stop words) are removed from the text in accordance with stop word list (includes 571 words) taken place in http://members.unine.ch/jacques.savoy/clef. Cleaned texts are recorded in the database.

Stemming

One of the important issues of the classification of the web pages is the feature extraction from the text. In this study, the words are used in the texts as the features and a dictionary is created. Stemming is determined by using Porter Stemming [11] algorithm. In this way, it is provided the words from similar originality to be represented with a word in the dictionary.

Pruning of the words

After identifying the stemming, data set was created. Rows show that the texts and columns show the words so the features. Data table can be created by writing the numbers (frequency) of related words on the table or by checking whether the words exit in the text. In this study, in order to reduce the number of features contained in the set of data with low frequency words in the data set was removed then examined whether or not the corresponding word is in the text and data set consisting of 0 and 1 values were obtained. Each row in the data set is a feature vector belongs to the text and is labeled depending on initial information whether it was related to Turkey.

K-NN classifier

K-NN algorithm is classification algorithm well known and widely used in many fields. A feature vector belongs to document which would be classified in K-NN algorithm is compared with training features vectors. It is used one of the criteria in comparison procedure such as Euclidian distance, cosine measure and so on. Tested document class is determined based on the classes of k training vector class with the best similarity. In this study, cosine was chosen as similarity measure and k as 1. Similarity measure takes the values between 0 and 1. 1 shows that the vectors are the same. Cosine measure was shown in Equation (1). Detailed information about K-NN algorithm can be found in Cover and Hart’s work [12].

\[
\text{sim}(d_i, q) = \sum_{j=1}^{w} W_{ij} \times W_{q,j} \\
\sqrt{\sum_{j=1}^{w} W_{i,j}^2} \times \sqrt{\sum_{j=1}^{w} W_{q,j}^2} = \frac{\sum_{j=1}^{w} W_{ij} \times W_{q,j}}{\sqrt{\sum_{j=1}^{w} W_{i,j}^2} \times \sqrt{\sum_{j=1}^{w} W_{q,j}^2}}
\]

(1)

\( W_{ij} \) is the weight of term in the text, \( d_i \) training vector and \( q \) class is the vector to have wished to be found.

RESULTS:

All experiments were carried out on our own data set. A dictionary was created by pruning character; stop words pruning, finding word stemming processes on the texts getting from the news channels RSS. 2321 words belong to 325 text are included in the dictionary. After to be created the dictionary, data table was realized by taking the
frequency of the words in the texts into consideration. If the frequency of the word is less than 2, it was removed from table so feature size was reduced (1386 words). Like the words TBMM, Çankaya which are considered to be identifier for the Turkey are added to the dictionary and so the creation of dictionary is completed.

Bitwise term weighting has been carried out taking into account the frequencies in the table, and the table has been made of 0-1s. Each row in the table is the one that includes the text, the text consisting of 0 and 1 values are expressed in terms of a feature vector. Each text in the data set table, whether or not belonging to Turkey was established as training set for the classifiers.

100 various texts were taken from international news sites disregarding their categories randomly in order to measure the performance of proposed method. Test texts are taken from news sources are shown in Table 1. A feature vector was created for each test text through the creation of data set by stepping similar.

Similarity values belong to all the vectors and taken place in training set to a feature vector of the test set were calculated by using cosine similarity. The class of the text was tried to be determined by using various k values for K-NN algorithm. The best results were obtained when k was selected 1. In this case, similarity values obtained for a test text belonging to the highest class test text has been recognized as a class. It has been observed that the proposed method is determined the text whether or not belongs to Turkey with 90% accuracy.

- THE DISTRIBUTION OF TEST TEXTS ACCORDING TO NEWS CHANNELS

<table>
<thead>
<tr>
<th>Distribution of test texts</th>
<th>Aljazeera</th>
<th>China org</th>
<th>CNN</th>
<th>IRNA</th>
<th>The Moscow times</th>
<th>Reuters</th>
<th>Spiegel Int.</th>
<th>UK World News</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of texts</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>8</td>
<td>6</td>
<td>20</td>
<td>13</td>
<td>27</td>
</tr>
</tbody>
</table>

CONCLUSION

In this study, an application was developed towards testing the texts about Turkey taken place in web pages of international news channels. For this purpose, the process of creation of dictionary and training data set were realized by using sample texts from 8 different channels. Then, whether the news texts from the news sites were BELONGING to Turkey was determined by using K-NN algorithm and cosine similarity. With recommended 150 seconds intervals, it has the ability of scanning RSS channels, determining new added news text, and if it is related to Turkey. In this way, the news channels, It can be monitored in the news RSS automatically. Situated on the Web on the subject of a lot of news stories can be measured with high accuracy is useful for the concerned one such as a quick news monitoring and time saving.

In future studies, system performance could be increased with decreasing of the words in the dictionary with different feature extraction algorithms, and using other classification algorithms in literature. The area of research and use can be increased by adding various language supports.
ACKNOWLEDGMENT

This study has been supported by Scientific Research Project of Selcuk University (Project No:13701497).

References


