International Journal of Computer Networks and Communications Security

VOL. 1, NO. 5, OCTOBER 2013, 189–194
Available online at: www.ijcncs.org
ISSN 2308-9830

A Study on Analysis of SMS Classification Using TF-IDF Weighting

Dr. Ghayda A. Al-Talib[1], Hind S. Hassan[2]

[12]Dept. of Computer Sciences, College of Mathematics and Computer sciences, University of Mosul, Mosul, Iraq.

E-mail: [1]ghaydatalib@yahoo.com, [2]hind_computers@yahoo.com

ABSTRACT

SMS classifying technology has important significance to assist people in dealing with SMS messages. Although sms classification can be performed with little or no effort by people, it still remains difficult for computers. Machine learning offers a promising approach to the design of algorithms for training computer programs to efficiently and accurately classify short text message data.. In this paper we introduce a weighting method based on statistical estimation of the importance of a word for an SMS categorization problem, which will classify Mobile SMS into predefined classes such as occasions, friendship, sales etc. All sms are converted into text documents. After preprocessing vector space model is prepared and weight is assigned to each term. This weighting method based on statistical estimation of the importance of a word for an SMS categorization problem. The experiments reported in the paper shows that this weighting method improves significantly the classification accuracy as measured on many categorization tasks.

Keywords: *Data Mining, text Classification, SMS, vector space model, TF-IDF Technique.*

1 INTRODUCTION

In the recent few years Short Message Service (SMS) has emerged as a popular means of communication between mobile users. The concept of SMS (Short Messaging Services) was highly successful, and soon became almost as important as the facility to have a voice communication. Today, this service is almost free, or being offered at a negligible cost [1].

Text Classification is the process of classifying documents into predefined classes based on its content. Text classification is important in many web applications like document indexing, document organization, spam filtering etc. [2].

In text classification, a text messages may partially match many categories. We need to find the best matching category for the text messages.

The term frequency-inverse document frequency (TF-IDF) approach is commonly used to weigh each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories [3].

In this paper, we use TF-IDF weighting model, which considers that if the term frequency is high and the term only appears in a little part of documents, then this term has a very good differentiate ability. This approach emphasizes the ability to differentiate different classes more, whereas it ignores the fact that the term that frequently appears in the documents belonging to the same class, can represent the characteristic of that class more[4].

We put forward the novel improved TF-IDF approach for text classification, and will focus on this approach in the remainder of this paper, and will describe in detail the motivation, methodology, and implementation of the improved TF-IDF approach.

2 SMS DOCUMENT RETRIEVAL

Mobile phone devices belonging to four categories have been stored in a local database for further processing. We made a small procedure through a program to convert them in to XML file as the following concern points:

1) Each message has verbs, nouns, adjectives and remaining sentence. In this we made our procedure to find them as the first step.

2) English dictionary, stop words list, previously have been used which are previously converted to XML file.

3) For building the database, all the messages have been entered in to the program and the result will be the repetition of each word in that message, as shown in Table 1.

4) Finally, the classes of all SMS messages as: occasion, greetings, friendships, sales categories, are assigned.

Table 1: The words acted in to xml file format

Occasions	
Words	Frequency
Year	139
Happiness	19
Prosperity	9
Others …	…

greetings	
words	Frequency
year	20
happiness	3
Prosperity	12
Others …	…

sales	
words	Frequency
year	24
happiness	1
Prosperity	15
Others …	…

Friendships	
Words	Frequency
Year	30
Happiness	11
Prosperity	1
Others …	…

3 SMS PROCESSING

This stage is crucial in determining the quality of the next stage, that is, the classification stage. It is important to select the significant keywords that carry the meaning, and neglect the words that do not contribute to the distinguishing between the documents; this stage consists of the following steps:

3.1 The replacement of the abbreviations

An abbreviation is a short way of writing a word or a phrase that could also be written out in full. Usually, but not always, it consists of a letter or group of letters taken from the word or phrase. For example, the word abbreviation can itself be represented by the abbreviation abbr, abbrv Or abbrev [5].

A collection of 1500 words have been collected with their abbreviations and stored in XML files, to replace all the existing abbreviations in messages with the original words that they mean.

3.2 Stop Word Removal

Many of the most frequently used words in English are useless in Information Retrieval (IR) and text mining. These words are called 'Stop words'. Stop-words, which are language-specific functional words, are frequent words that carry no information (i.e., pronouns, prepositions, conjunctions). In English language, there are about 400- 500 Stop words. Examples of such words include ('the', 'of', 'and', 'to'). We need to remove these Stop words, which has proven as very important step because it reduces the size of text to be processed [6].

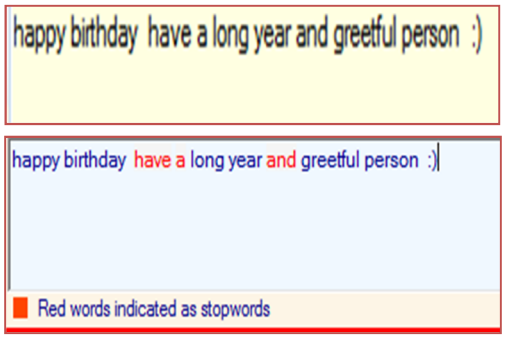

Fig. 1. The normal message before and after stop words process

3.3 Part of speech tagging

Part-of-speech (POS) tagging is the task of determining the correct parts of speech for a sequence of words. POS tagging is useful for a large number of applications: It is the first analysis step in many syntactic parsers. it is used in information extraction, speech synthesis, lexicographic research, term extraction, and many other applications [7].

Every term in the document has a part of speech tag such as noun, verb, adjective and adverb. As human, we can see that not all the word forms contribute to the meaning of a document in the same amount. For example it is expected that adverbs are kind of transition words and do not tell much about the content in the document, whereas nouns tells much more [8]. In this research we have used POS tagging to extract the verbs, adjectives and names as features and neglect the other parts of speech.

Fig. 2. The tagging words have different colors in each sentence classified

3.4 Stemming technique

Stemming techniques are used to find out the root/stem of a word. Stemming converts words to their stems, which incorporates a great deal of language dependent linguistic knowledge. Behind stemming, the hypothesis is that words with the same stem or word root mostly describe same or relatively close concepts in text and so words can be conflated by using stems. For example, the words, user, users, used, using all can be stemmed to the word 'USE'. In the present work, the Porter Stemmer algorithm [9], which is the most commonly used algorithm in English, was used.

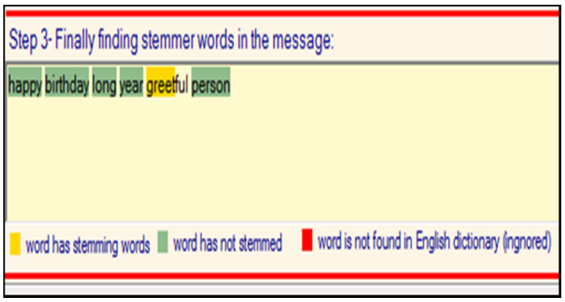

Fig. 3. The stemmed words have yellow color in each sentence

4 CREATING DICTIONARIES

There is a dictionary for each type of SMS have been created which contain the words that are extracted from each type with the number of occurrence in each document, and it is stored in XML file, then an English dictionary have been used in order to search for each word in the SMS and if it is not found there it is considered as a foreign word and it will be neglected.

5 VECTOR SPACE MODEL

The vector space model defines documents as vectors (points) in a multidimensional space where the axes (dimensions) are represented by terms. Depending on the type of vector components (coordinates), there are three basic versions for this representation: Boolean, term frequency (TF), and term frequency_ inverse document frequency (TF-IDF) [10] , which is used in this research.

6 TF-IDF TECHNIQUE

TF-IDF is evolved from IDF which is proposed by Sparck Jones with the heuristic intuition that a query term which occurs in many documents is not a good discriminator, and should be given less weight than one which occurs in few documents [11].

The formula of TF-IDF is:

$$TF\text{-}IDF(ti,dj)=tf(ti,dj)\log N/ni \quad (1)$$

Where tf(ti,dj) represents the term frequency of

term i in document j, N represents the total number of documents in the dataset, represents the number of documents where the term i appears [12].

The basis of TF*IDF is from the theory of language modeling that the terms in a given document can be divided into (with and without) the property of eliteness [11], i.e., the term is about the topic of the given document or not. The eliteness of a term for a given document can be evaluated by TF and IDF which is used for the measure of importance of this term in the collection.

However, there are some deficiencies of TF*IDF method. The first one is that it is sometimes criticized as 'ad-hoc' because it is not directly derived from a mathematical model of term distribution or relevancy analysis although usually it is explained by Shannon's information theory [13], The second one is the dimensionality of text data which affect the size of the vocabulary across the entire data-set. And it brings out a huge computation of the weight of each term occurring in each document [14].

7 RESULT AND DISCUSSION

In this study we analyzed 4 categories of sms these are sales, occasion, friendship, and greeting sms, we randomly sampled two-thirds of the sms for training and used the remaining one-third for testing. Normally, as an example, we have sample message as it appearing:

"Thank you for your May 15 telephone order for 475 TV/VCR coaxial cables. Delivery of our catalog items generally takes less than a week. Larger orders such as yours may take two to three weeks. We are pleased to notify you, however, that your large order qualifies you for our new 20% bulk discount, applied to all orders over $200. (As you will see on the accompanying invoice, we have already deducted your discount from the total price of your order.) " .

The program will omit unwanted or repeated words in the sentence by using the stop words list; the words are referred by red color in order to be ignored by the next step as shown in the Figure 4.

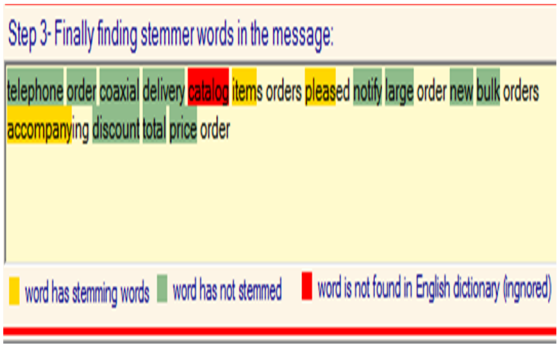

Fig. 4. Stop word removal

The message is still long and it is difficult to be recognized and classified. A new technology has been proposed by using tagging words to keep some specified kinds of words more likely verbs, adjective and nouns. A library named by OpenNLP has been used to keep eyes on the criteria words only and ignore other kinds of sentence's words. Figure 5 shows the tagging words, where tokens are mentioned by different colors in the same figure. We colorized the verb by green light color while nouns in blue and for the adjectives are in orchid color. The black color will be ignored in the next step from the message.

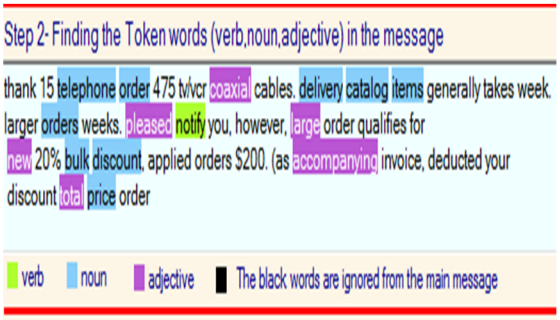

Fig. 5. The tokens in the message

The final step is stemming words, by using some more technique to avoid the spelling correction. In fact an English dictionary was built by collecting over 5800 English words then a comparison of the stemmed words with this dictionary was done. Figure 6 shows the result of the steaming technique.

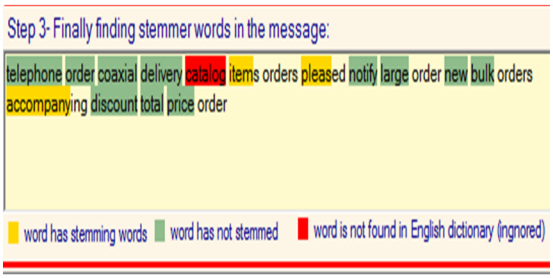

Fig. 6. The stemmed words in the message

As a result the following words will remain after the above processing: "telephone, order, coaxial, delivery, pleas, large, new, bulk, accompany, total price" In the last steps, what we need now is to get the weights of the words from the XML file format.

The words that are selecting for message classification are the following "telephone, order, coaxial, delivery, pleas large new, bulk, accompany, total price". By applying TF-IDF technique on the remaining weighted words, the output will be as in Figure 7.

No.	Word	Found in EN Dictionary	Occations [tf-idf]	Greetings [tf-idf]	Friendships [tf-idf]	Seales [tf-idf]	[Inspecting Type]
0001	telephone	True	0	0	0	0.115577889447236	Seales
0002	order	True	0.0181818181818182	0.0181818181818182	0.0181818181818182	1.18592964824121	Seales
0003	coaxial	True	1.18592964824121	1.18592964824121	1.18592964824121	0.0050251256281407	Seales
0004	delivery	True	0.0181818181818182	0.010752688172043	0.010752688172043	0.116142106305969	Seales
0005	plea	True	0.116142106305969	0.116142106305969	0.116142106305969	0.116142106305969	Seales
0006	large	True	0.116142106305969	0.116142106305969	0.116142106305969	0.125628140703518	Seales
0007	new	True	1.29090909090909	0.505376344086022	0.0912927418889577	1	OCCATION
0008	bulk	True	1	1	1	0.0100502512562814	Seales
0009	total	True	0.0100502512562814	0.0100502512562814	0.0161290322580645	0.0435510887772194	Seales
0010	price	True	0.0435510887772194	0.0435510887772194	0.0435510887772194	1.08542713567839	Seales

Fig. 7. The final result of the TF-IDF technique

The decision in majority goes to sales by 99% according to the whole words except new which is 1% percent goes to occasion's type. The word which has a high frequency on those documents will willing that kind of that document.

8 CONCLUSION

Text categorization is a hot research topic in current information retrieval, and is an important branch of data mining and information retrieval. How to improve the classification accuracy is an important topic in text categorization, in order to solve this problem, much research has been done to find new classifiers which will improve the accuracy, whereas this paper tries to improve the accuracy by proposing an improvement on TF-IDF weighting method. From the experiments, we can see this improvement increases the accuracy significantly, therefore we think this improvement is promising.

9 REFERENCE

[1] SHILPA MEHTA, U ERANNA, K. SOUNDARARAJAN, A Neural Technique for SMS Classification Using Keywords Search and Identification of Captured Messages, Using Hebbian Learning, International Journal of

[2] Engineering Sciences Research-IJESR, Vol. 03, No. 03, July 2012.

[3] Deepshikha Patel, Monika Bhatnagar, Mobile SMS Classification, International Journal of Soft Computing and Engineering (IJSCE), Volume-I, Issue-I, March 2011.

[4] Mingyong Liu and Jiangang Yang, An improvement of TF-IDF weighting in text

categorization, International Conference on Computer Technology and Science, Vol. 47,No 9, 2012.

[5] ZHANG Yun-tao, GONG Ling, WANG Yong-cheng, An improved TF-IDF approach for text classification*, Journal of Zhejiang University SCIENCE, ISSN 1009-3095, 2005.

[6] https://en.wikipedia.org/wiki/Abbreviation.

[7] V. Srividhya, R. Anitha, Evaluating Preprocessing Techniques in Text Categorization, International Journal of Computer Science and Application Issue, ISSN 0974-0767, 2010.

[8] Sandipan Dandapat, Part-of-Speech Tagging for Bengali, Department of Computer Science and Engineering Indian Institute of Technology, Kharagpur January, 2009.

[9] Kerem c_ elik, a comprehensive analysis of using wordnet, part-of-speech tagging, and word sense disambiguation in textcategorization, b.s., computer engineering, bah_ce_sehir university, 2009.

[10] Willett, P, The Porter stemming algorithm, electronic library and information systems, Vol. 40, 2006.

[11] Marhov Z. , and larose D.T. , Data mining the web, wiley, 2007.

[12] Man and Cybernetics, TF-IDF, LSI and Multi-word in Information Retrieval and Text Categorization, IEEE International Conference on Systems,2008 .

[13] DEQING WANG AND HUI ZHANG, Inverse-Category-Frequency Based Supervised TermWeighting Schemes for Text Categorization*, JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 29, 209-225, 2013

[14] Thomson Avenue, Understanding Inverse Document Frequency On theoretical arguments for IDF, Journal of Documentation 60 no. 5, 2004.

[15] D. M. Christopher and S.Hinrich, Foundations of Statistical naturallanguage processing. MIT Press. Cambridge, Massachusetts, 2001.