Semantic Approach to Identify Topics from Product Reviews

Deepthi Devaiah D, Hithaishy B J, Pallavi Bhat
Department of Computer Science and Engineering
Sri Jayachamarajendra College of Engineering, Mysore, Karnataka, India

Dr. Anil Kumar K M
Department of Computer Science and Engineering
Sri Jayachamarajendra College of Engineering, Mysore, Karnataka, India

Abstract- Topic identification is the construction of useful labels for set of documents. It is essential in connection with categorizing search applications, where several sets of documents are delivered and an expressive description for each category must be constructed on the fly [1]. The main objective of topical identification is to assign topics for each of the text file. This technique has been illustrated by different methods on the basis of term frequency. In order to increase the efficiency, synonyms of the words occurring maximum number of times is taken into consideration. Since most of the topics are nouns, we came up with an idea to consider only nouns from the text files and this is done using parts of speech tagger. Since it is not always necessary that the maximum occurring word matches with the topic, another method called base line with variation is used. The topics for each set of text files are evaluated and kappa value is computed. After annotation, all the above mentioned approaches are repeated again and new results are obtained. The synonyms approach is found to be effective and provides efficient results on the data set.

Keywords – Semantic Approach, Product Reviews, Topic Identification

I. Introduction

Text mining, also known as text data mining or knowledge discovery from textual databases refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases. Text mining focuses on extracting a small amount of information from text with high reliability [2]. Our main objective is to identify the topic of a set of product reviews.

Example: **"*Love my new g3, just received this camera two days ago and already love the features it has. Takes excellent photos. The camera is easy to use once you get used to it. However, using the lcd seems to eliminate this minor problem. Overall it is the best camera on the market and I give it 10 stars!*"**

Based on the above example we can conclude that the information is about camera. This concluding information is present in the first and last sentence also. We cannot determine the subject content, if we haven't identified the topic. The product reviews were collected from various websites of the World Wide Web and no attempt was made to correct any grammatical or logical mistakes in them. Because of topical identification it becomes easy to know the content of the text file, else it is extremely hard and time consuming. The remainder of the paper is organized as follows. In section II, Proposed Algorithm is discussed. Experimental results are presented in section III. Concluding remarks are given in section IV.

II. Proposed Algorithm

We collected text files on reviews of different products. The texts were collected manually from website like Wikipedia. We took reviews of 4 different products such as Apex DVD Player, Creative Zen Player, Nokia 6610 and Canon Camera. In general, for all the approaches we first try to eliminate symbols that do not contribute to determine the topic. First, we eliminate all the punctuation marks in the text files by getting each character from the file and comparing it with the range of ASCII values for symbols that is only characters in the range [A-Za-z] were checked. This file is considered for further processing.

Given that the text file is of particular topic, one would expect particular words to occur more or less frequently. "Camera" will appear more often in documents about camera, and "the" and "is" will appear equally. Since these words such as "the" and "is" cannot be the topic, these should be removed before undergoing further processing. The text file is then freed from these words called stop words [7]. Stop words are prepositions and pre-nouns which frequently occur in a text file, and the number of such occurrences is more than the number of times the keywords appear in the file. A few examples of stop words are listed in Table 1. For example, consider a line from the text file based on camera review, ***"This is one of the best cameras I have purchased so far***". After reading the characters in the acceptable range and eliminating the stop words, we obtain a line of the form ***"cameras purchased***".

Table -1 List of Stop Words

a	**certain**	**keep**
across	**else**	**look**
and	**for**	**my**

The remaining words in the text file are processed. The word which has maximum frequency is compared with the topic. This topic is assumed for each text file. If the word with maximum frequency matches with the topic of the text file, then it is taken as match else mismatch. Accuracy is evaluated for a set of text files. This is called as **base line approach**.

It is not always necessary that the maximum occurring word should match with the topic. So in order to get more accurate results, we use window size concept. In this concept, the words with maximum frequency are put into window slot 1. The first maximum and second maximum occurring words are put in window slot 2 and so on. Accuracy is measured for each window size for Set 1 of product reviews. Initially the accuracy increases and at a certain point accuracy becomes constant and we fix it as the threshold value. The corresponding window size is the threshold. The testing is done for further set of text files by considering the window size for which the threshold is obtained. This is called as **base line with variation approach**. Usually introduction and conclusion of the text file contains the topic. So with this logic, instead of checking the word with the maximum frequency in the whole text file, we consider the first line and last line of a text file and determine the word which occurs maximum number of times, compare with the topic and check if it is match or mismatch. This is called as **first and last line approach**.

In order to further improvise the above mentioned approaches, instead of considering only the topic we consider synonyms of the topic. WordNet [8] can be used to find the synonym set of the possible words that may form the topic of the text file to give an overview of the text file. We have identified synonym sets as shown in Table 2. While finding the synonym of the topic, we have considered only Unigrams that is if the topic is DVD player, then the synonym of DVD and player are found separately using WordNet and taken for further processing. The word which has maximum frequency in the text file is now compared with the topic and its synonyms, if it matches then it is considered as a match else mismatch. This is called as **synonym approach**. In Table 2 we have taken relevant synonyms from WordNet and we have not taken brand names of product into consideration.

Table -2 List of Synonyms

Apex DVD player	**DVD player, optical disk, optical disc**
Creative Zen	**Player, ipod**
Nokia 6610	**Phone, mobile**
Canon, Nikon camera	**Camera**

Usually most of the topics of the text files are nouns. So, instead of considering all the words in the text file for processing, only the nouns are considered. In the **parts of speech tagger approach**, each text file is subjected to an online parts of speech tagger[9] which tags every word in this text file as a noun or verb or preposition or pro-noun and so on. Only nouns from this newly tagged text file are taken in another file and it is now subjected to maximum frequency approach and first-last line approach. For example, consider a line from the text file based on camera review, ***"Powershot g3 is the flagship of canon's powershot series and it is an slr-like camera"***. After reading the characters in the acceptable range, eliminating the stop words and applying online speech tagger we obtain a line of the form ***" NN powershot NN g3 VBZ is DT the NN flagship IN of NN canon POS's NN powershot NN series CC and PRP it VBZ is DT and JJ slr-like NN camera"***. The above procedure is carried out for multiple files using batch files (run.bat). Based on the expected output, the accuracy is calculated. Few parts of speech tagger is defined in Table 3.

Table -3 Parts of Speech Tagger

NN – Singular Noun	POS – Possesive ending
NNS – Plural Noun	VBZ – Verb
NNPS – Proper plural noun	IN – Preposition

Another **approach is parts of speech tagger using synonyms.** In this the maximum occurring word (noun) in the tagged file is compared with all the synonyms of the topic obtained using WordNet and then accuracy is evaluated. The same approach is applied to determine the first and last line of the text file. All the above methods were done before evaluating annotations. Annotations can be evaluated using the kappa value. In order to calculate the kappa value, the text files were given to set of people. They read the text files and annotate each text file. By comparing the annotations given by the set of people the kappa value can be calculated using Equation 1.

kappa= (1)

where,

p_{pos} : the positive agreement. p_o : the overall proportion of agreement. p_{neg} : the negative agreement. p_e : the agreement expected by chance. kappa : the measure of agreement corrected by chance.

These measures can be best explained through an example as shown in Table 4. Suppose the annotators A and B want to find the agreement of the topic DVD player among a set of text files. 97 text files are taken for consideration.

Table -4 Example for Annotator Agreement

Annotator 2

Annotator 1	Positives	Negatives	Total
Positives	75	5	80
Negatives	3	14	17
Total	78	19	97

- p_{pos} : The number of positives that both annotators agree, divided by the number of all positives for both annotators. p_{pos}= =0.94
- p_{neg} : It is calculated in a similar way. $p_{neg}=\frac{14+14}{19+17}=0.77$
- p_o: It is calculated by the positives both readers agree on plus the number of negatives both readers agree on, divided by the total number of text files. p_o= =0.91
- The joint agreement expected by chance (p_e) is calculated for each combination.

  p_o= =0.697
- kappa: It is calculated by subtracting the portion of the annotations which are expected to agree by chance from the overall agreement, and dividing the remainder by the number of cases on which agreement is not expected to occur by chance[10]. kappa= =0.70

Table -5 Guideline for the strength of the kappa value

kappa value	Strength of Agreement
<0	Poor
0-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

The agreement for the topic is substantial. In the paper by Landis[11] the guidelines for the strength of the kappa values is given as shown in Table 5. All the above mentioned approaches are again applied on the set of text files which the annotators have agreed upon. Using these methodologies, the accuracy A, for each set of review files, 'i' is calculated using the Equation 2.

A(i)= (2)

where,

X(i) : Number of files which agree upon the topic
Y(i) : Total number of text files
The same formula is used to determine the accuracy for all the approaches.

III. Experiment and Result

Our objective is to identify the appropriate topic from the set of product reviews taken for analysis. In the preprocessing, the punctuations and stop words are eliminated from these text documents and they are further processed. On these text documents different approaches are employed such as, **base line method, synonym method, parts of speech tagger method, parts of speech tagger with synonyms method and baseline with variation**. The four sets of product reviews considered are of DVD player, Creative Zen player, Nokia phone and Canon camera. The results of base line method are tabulated in Table 6.

Table -6 Results of Base Line method

Approach / Topic	DVD Player	Zen Player	Nokia phone	Canon camera	Average accuracy
Maximum occurrence	**37.11**	**29.27**	**85.29**	**63.29**	**53.79**
First and Last line	**27.84**	**21.05**	**61.76**	**62.03**	**43.16**
First Line	**17.52**	**14.73**	**58.82**	**60.75**	**37.85**
Last Line	**21.64**	**12.63**	**41.17**	**36.70**	**28.04**

Table -7 Results of Synonyms method

Approach / Topic	DVD Player	Zen Player	Nokia phone	Canon camera	Average accuracy
Maximum occurrence	**42.26**	**31.57**	**85.29**	**63.29**	**55.60**
First and Last line	**35.05**	**25.26**	**61.76**	**62.03**	**46.02**
First Line	**39.17**	**15.46**	**55.88**	**62.02**	**43.13**
Last Line	**27.83**	**25.26**	**58.82**	**41.77**	**38.34**

While the first method that is, base line method played a significant role in topical identification of the text documents, it can be noted that the synonyms approach shows an improvement in the accuracy, as shown in Table 7. We observe that the accuracy of classification for maximum occurrence and first and last line approach has increased when compared to the base line method. The third approach that is, parts of speech tagger method is an idea proposed in order to get better results. The accuracy results in Table 8 show that this method is better than base line method, but not better than synonyms method. The fourth method combines both parts of speech tagger and synonyms method. In comparison of this method with the synonyms method, there is an improvement in the accuracy Table 9. Due to the page limit mentioned, we are not able to include all the graphs and the details of all the experiments.

Table -8 Results of Parts of Speech Tagger method

Approach / Topic	DVD Player	Zen Player	Nokia phone	Canon camera	Average accuracy
Maximum Occurrence	**22.68**	**31.58**	**85.29**	**62.02**	**50.39**
First and Last line	**24.74**	**20.00**	**61.76**	**64.55**	**42.75**
First Line	**12.37**	**12.63**	**50.00**	**62.02**	**34.26**
Last Line	**20.62**	**12.63**	**35.29**	**34.18**	**25.68**

Table -9 Results of Parts of Speech Tagger with Synonyms method

Approach / Topic	DVD Player	Zen Player	Nokia phone	Canon camera	Average accuracy
Maximum Occurrence	**22.68**	**33.68**	**85.29**	**64.56**	**51.55**
First and Last line	**27.87**	**24.21**	**67.65**	**67.09**	**46.69**
First Line	**14.43**	**12.63**	**55.88**	**62.03**	**36.24**
Last Line	**22.68**	**18.95**	**44.12**	**35.44**	**30.30**

All the above specified results were obtained before annotations. The text documents are subjected to topic evaluation. After this the annotation agreement is checked by evaluating the kappa value. If the kappa value is ~1 then only the text files which are in agreement with the topic are considered and the text files are then subjected to the above mentioned approaches. Figure 1 shows the comparison of the accuracies before and after annotations of Canon Camera text files in Base line method. Figure 2 shows the comparison of the accuracies before and after annotations of Canon Camera text files in Synonyms method. Figure 3 shows the comparison of the accuracies before and after annotations of Canon Camera text files in Parts of Speech Tagger method. Figure 4 shows the comparison of the accuracies before and after annotations of Canon Camera text files in Parts of Speech Tagger with Synonyms method.

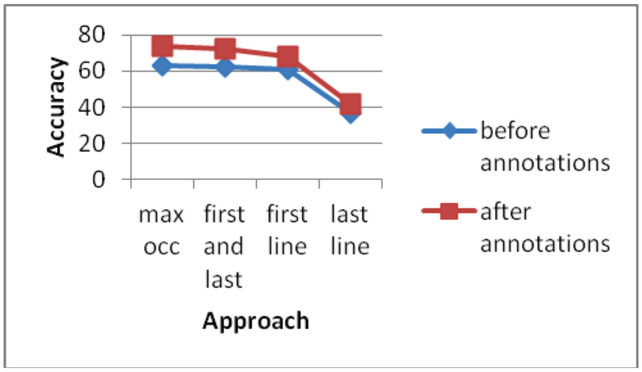

Figure 1. Base Line method

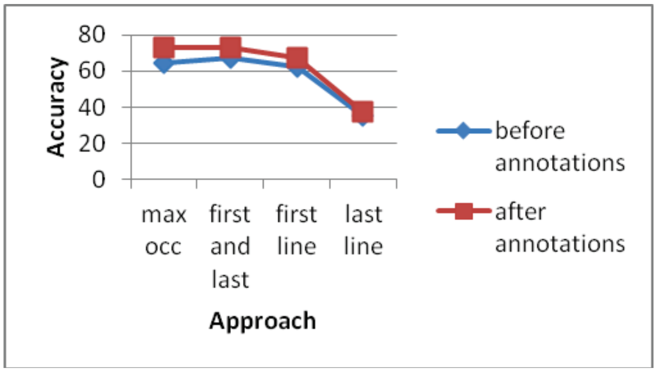

Figure 2 . Synonym method

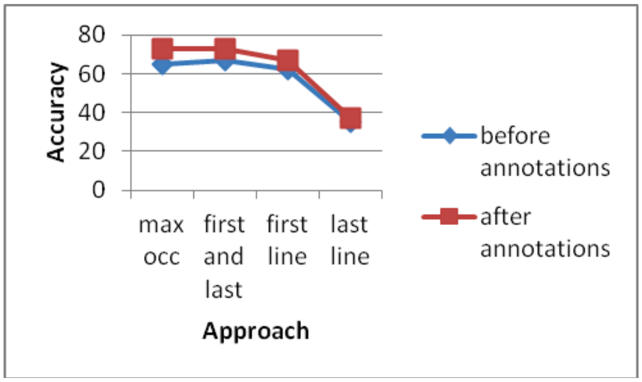

Figure 3. Parts of Speech Tagger method

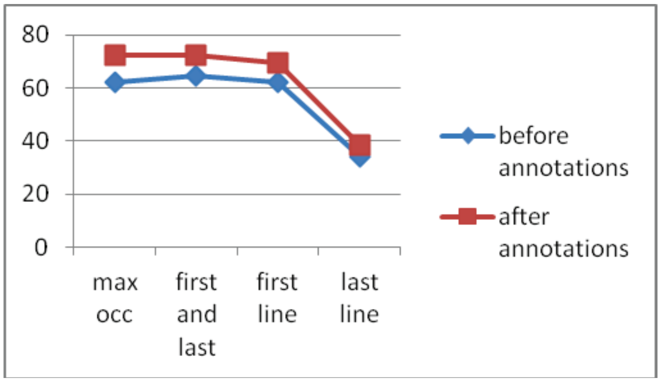

Figure 4. Parts of Speech Tagger with Synonym method

Table 10 shows the accuracy results of baseline with variation method which is the another variation of Baseline method. Threshold is obtained at window slot 5. The accuracy for other set of text files are evaluated with window slot 5.

Table -10 Baseline with Variation

Approach Topic	DVD player	Zen player	Nokia phone	Canon camera
window slot 1	37.113	-	-	-
window slot 2	61.856	-	-	-
window slot 3	73.195	-	-	-
window slot 4	80.412	-	-	-
window slot 5	81.443	64.210	94.117	96.202
window slot 6	81.443	-	-	-

IV.CONCLUSION

We have experimented with different approaches to easily and efficiently detect the topic of the product reviews. We have done the analysis of these approaches on nearly 300 product review files. We found that synonym method performs better than the base line and parts of speech tagger methods. When we considered the same approaches for first line, last line, both first and last line we found that our conclusion based on both first line and last line is more accurate than separately considering first and last lines. We have also found that the experiments done on the product reviews after annotation were better than experiments done before annotations. Also, we found that synonyms method, with an efficiency of 60% performs better than our approach using parts of speech tagger, with an efficiency of 52%, for the entire document after annotations. Future enhancement can be done by using topic identification in opinion mining to find the topic or product on which the user puts forward his reviews.

REFERENCES

[1] Topic identification- Framework and Application(2004) by Benno Stein, Sven Meyer Zu Eissen

[2] Text Mining: The state of the art and the challenges, Ah-Hwee tan, Ken Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore-119613

[3] A paper by Manu Aery, Naveen Ramamurthy, Y. Alp Aslandogan ,Department of Computer Science and Engineering , The University of Texas at Arlington

[4] Topic Identication Using Wikipedia Graph Centrality by Kino Coursey, University of North Texas and Daxtron Laboratories, Inc,kino@daxtron.com

[5] Topic Identification: Framework and Application, Benno Stein(Paderborn University, Germany, stein@upb.de) Sven Meyer zuEissen(Paderborn University, Germany,smze@upb.de).

[6] Topic Identification in Discourse by Kuang-hua Chen, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, R.O.C(khchen@nlg.csie.ntu.edu.tw).

[7] http://www.ranks.nl/resources/stopwords.html

[8] WordNet – A lexical database for English from Princeton University.

[9] Parts of Speech Tagging – Cognitive Computation Group. Cogcomp.cs.illinois.edu

[10] http://igitur-archive.library.uu.nl/dissertations/2006-1011-200545/c9.pdf

[11] http://www.stattutorials.com/SPSS/TUTORIAL-SPSS-Interrater-Reliability-Kappa.htm