International Journal for Modern Trends in Science and Technology
Volume: 03, Issue No: 09, September 2017
ISSN: 2455-3778
http://www.ijmtst.com

Feature Extraction of Customer Reviews Using Frequent Pattern Mining Algorithm

S.Anitha[1] | Dr.K.Karpagam[2]

[1]M.Phil Scholar in Department of Computer Science, Bharathidasan University, Pudukkottai, Tamilnadu, India.
[2]Professor, Department of Computer Science, Bharathidasan University, Pudukkottai, Tamilnadu, India.

To Cite this Article
S.Anitha and Dr.K.Karpagam, "Feature Extraction of Customer Reviews Using Frequent Pattern Mining Algorithm", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 09, September 2017, pp.-91-95.

ABSTRACT

Selling the product through the Web has become more popular because of online shopping. As e-commerce is becoming more and more familiar, the number of customer reviews that a product receives grows quickly. For a accepted product, the amount of reviews can be in hundreds or level thousands. This makes it difficult for a potential customer to read them in order to make a decision to buy the product. The main objective of this work is to discuss about developing an information extraction system which mines customer reviews in order to build a model to extract important product feature and their evaluation by reviewers. In this paper, we present a frequent pattern mining algorithm to mine a number of reviews and extract product features. Our new result indicates the algorithm outperforms the old pattern mining techniques used by previous researchers.

Keywords: *Association rule, Pattern mining, Product feature extraction, Text mining.*

I. INTRODUCTION

The Product feature extraction is important task of review mining and summarization. Opinion features are mined from product reviews based on data mining and natural language processing methods. A feature-based summary [1] of a large number of customer reviews of a product sold online is obtained. This issue will become increasingly important as more people are buying and expressing their opinions on the Web. This work tried to find out better product features from customer reviews. Opinion mining or sentiment analysis aims to determine whether the review sentences deliver a positive, negative or neutral orientation. Product feature extraction is critical to sentiment analysis, because the opinion orientation identification is significantly affected by the target features.

In order to find the frequent features, association mining is used. In this context, an item set is a set of words or a phrase that occurs together. The idea behind this technique is that features that appear on many opinions have more chance to be relevant, and therefore, more likely to be actually a real product feature. To mine frequent occurring phrases, each piece of information extracted above is stored in a dataset called a transaction set/file. Then it runs the association rule miner, which is based on the Apriori algorithm. It finds all frequent itemsets in the transaction file. Each resulting frequent item set is a possible feature. In this work, we define an item set as frequent if it appears in more than minimum support of the review sentences.

Frequent Features are stored to the feature set for further processing. This work used a much simpler mechanism and yet very effective. There

are also some uninteresting and redundant ones. Feature pruning aims to remove these incorrect features. Two types of pruning are presented: one is Compactness pruning that checks features that contain at least two words, which are named feature phrases [2], and removes those that are likely to be meaningless. Second is Redundancy pruning that removes redundant features that contain single words. The basic idea of feature based opinion mining is to determine the sentiments or opinions that are expressed on different features or aspects of entities. When text is classified at document level or sentence level it might not tell what the opinion holder likes or dislikes. If a document is positive on an object it clearly does not mean that the opinion holder will hold positive opinions about all the aspects or features of the object

II. RELATED WORK

Popesco[1] proposed OPINE for extracting components and attributes of the products reviewed by the consumers. They compute the point wise mutual information (PMI) between noun phrases and a set of metonymy discriminators associated with the product class. Their approach is based on the hypothesis that features associated with their product category tend to co-occur in reviews.

This work is closely related to Hue and Liu's [6] Work in on extracting product features from reviews. Using association mining they looked for the features that have been talked about by the people frequently. Based on the observation that features are generally nouns or noun phrases, they ran Apriori algorithm on the transaction set of noun/noun phrases to generate frequent itemsets. After producing candidate features they applied compactness pruning and redundancy pruning to remove those features that are not genuine. However, their proposed method was effective in discovering frequent features, but using Apriori leads to increase the execution time while dealing with large databases.

Chin-Ping[8] extended the above study by adding an additional step to prune possible non product features and opinion-irrelevant product features. They collect a list of positive and negative words from the general inquirer to determine the subjectivity of a review sentence. Then those frequent features which never or rarely co-occur with any positive or negative adjectives in review sentences are considered as opinion-irrelevant features and removed. We applied different pattern mining algorithm to enhance the precision and performance of the system simultaneously.

III. METHODOLOGY

Specific instances which are aim to extract from given dataset are defined as following

Definition 1: Product Feature

Product features refer to all the components, qualities or physical characteristics of a product such as size, color, weight, speed, etc.

Definition 2: Opinion Sentence

An opinion sentence is a sentence that consists of a least one product feature and its corresponding opinion word. An opinion word is a term used to reference a word that usually qualifies an object or an attribute of this object. They are usually adjectives and adverbs, but they can also be nouns and verbs. Simply [3], If a sentence contains one or more product features and one or more opinion words, then the sentence is called an opinion sentence.

The sentences "I bought this camera last year. Since then I have been very happy with its image quality". Here, the first sentence will be discarded and will be not further analyzed as no opinion word is found. The second sentence satisfies the definition of an opinion sentence as happy is a opinion word and image quality is a camera feature.

Definition 3: explicit and implicit feature

When a feature f is readibly available in a review R, f is called an explicit feature. There are cases where a feature f is not readibly available, in R, therefore it is considered to be an implicit feature. An explicit feature is a feature of a product which is directly talked about in review sentence. An implicit feature is a feature that is not explicitly mentioned in the sentence and it can be implied.

Example 1:

I. "The battery life of this camera is too short"

II. "This camera is too large"

In the first sentence, battery life is an explicit feature, while in the second one, size is an implicit feature. Size is not mentioned in this sentence, but it is easy to realize that large indicates a negative feature of the size attribute.

Definition 4: frequent and infrequent feature

A feature *f* is frequent if it appears in majority of the review sentences. *f* is called infrequent if it is only appeared in a few number of reviews. After putting all these definitions together we go through with general problem of identifying features in the reviews.

Most current researches focus on discovering

explicit product features. Generally, the current approaches are either supervised or unsupervised. Although, supervised approaches sound to be more accurate, but they need training set that is generated by the human. This approach is effective when the documents are not too away in terms of the subjectivity.

IV. PROPOSED TECHNIQUE

The architectural overview of our feature extraction system is given in Figure 4.1 and each system component is detailed subsequently. The system input is a product review dataset including a large number of reviews on products.

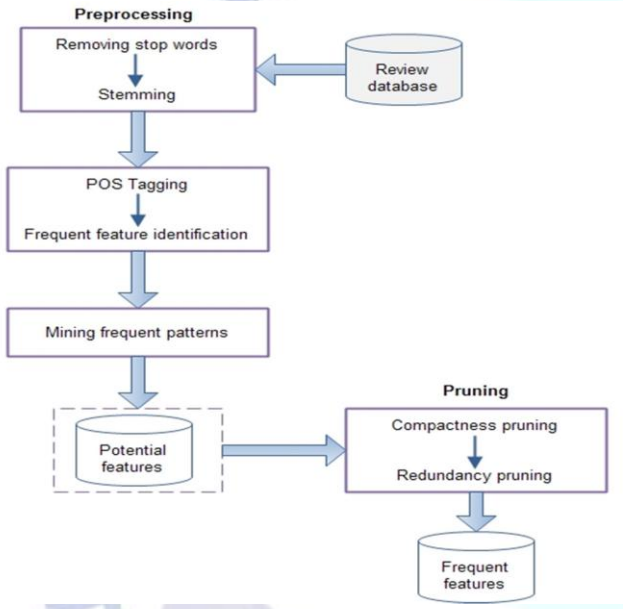

Fig.4.1. The System Framework

4.1 Phase 1: Preprocessing

In this work we perform some pre-processing of words including removal of stop words and stemming before going through the next steps.

4.2 Phase 2: Part-of-Speech-Tagging

The tagged sentences produced by the NLProcessor in this step, will play a very important role for the rest of the system. In feature identification [4], a data mining system will depend on the noun or noun phrases generated on this step to produce a number of frequent features. Also, the classification of sentiment will depend on the words classified both as adjectives and adverbs in this step to produce a set of possible opinion words.

As the only focused part of the sentences in our work is nouns or noun phrases, we apply a Part-Of-Speech tagger that we developed in PHP to identify the role of the words within the sentences.

4.3 Phase 3: Frequent Feature Identification

This sub-step identifies product features on which many people their opinions. Before discussing frequent feature identification [5], we first give some example sentences from some reviews to describe what kinds of opinions that we will be handling. Since our system aims to find what people like and dislike about a given product, how to find the product features that people talk about is the crucial step. However, due to the difficulty of natural language understanding, some types of sentences are hard to deal with. Let us see an easy and a hard sentence from the reviews of a digital camera:

"The pictures are very clear."

In this sentence, the user is satisfied with the picture quality of the camera, *picture* is the feature that the user talks about. While the feature of this sentence is explicitly mentioned in the sentence, some features are implicit and hard to find. For example,

"While light, it will not easily fit in pockets."

This customer is talking about the *size* of the camera, but the word *size* does not appear in the sentence. In this work, we focus on finding features that appear explicitly as nouns or noun phrases in

The idea behind this technique is that features that appear on many opinions have more chance to be relevant, and therefore, more likely to be actually a real product feature. The Apriori algorithm was used to generate the set of frequent itemsets. However, for this task there was no need of finding association rules among items, therefore only the part of the algorithm that finds frequent itemsets was interesting for these works.

4.4 Phase 4: Pruning

Association mining algorithms does not consider the position of the items in a given transaction. Thus [7], after running the algorithm on a sequence of words as an input transaction, it generates a number of candidates that may not be genuine features.

However, not all candidate frequent features generated by association mining are genuine features. Two types of pruning are used to remove those unlikely features.

Compactness pruning:

This method checks features that contain at least two words, which we call *feature phrases*, and remove those that are likely to be meaningless. The association mining algorithm does not consider the position of an item in a sentence. However [9], in a sentence, words that appear together in a specific

order are more likely to be meaningful phrases. Therefore, some of the frequent feature phrases generated by association mining may not be genuine features. Compactness pruning aims to prune those candidate features whose words do not appear together in a specific order.

Redundancy pruning:

In this step, we focus on removing redundant features that contain single words. To describe the meaning of redundant features, we use the concept of *p-support* (*pure support*). *p-support* of feature *ftr* is the number of sentences that *ftr* appears in as a noun or noun phrase, and these sentences must contain no feature phrase that is a superset of *ftr*. We use a minimum p-support value to prune those redundant.

V. EXPERIMENTAL RESULTS AND EVALUATION

To evaluate the effectiveness of the sentiment classification algorithm, the orientation associated with each feature in a sentence was analyzed manually in order to achieve a high degree of confidence. A correctly classified opinion is either a negative or positive opinion for a given feature, which was correctly identified by the system. All the features were identified automatically, and only the sentences with real features (frequent features) were analyzed. Sentences with candidate features were discarded.

Table 7.1: Sample data for evaluation tests

Product	Number of Opinions
Ipod Touch 8GB ;	135
Nikon D5000 86	76
Nikon P90 52	58
Xbox 360	51

Table 7.2: Effectiveness of sentiment classification

Product	Correctly Classified Opinions	Opinion sentences
Ipod Touch8GB	79%	593
Nikon D5000	93%	432
Nikon P90	54%	393
Xbox 360	63%	321

Sentiment classification effectiveness

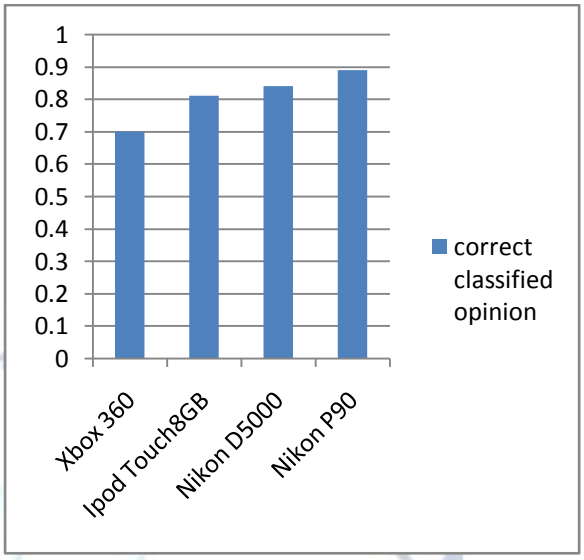

Fig 7.1: Effectiveness of opinion sentiment classification

The evaluation tests to address performance were made. A total of 120 opinions for Nikon D5000 stored in the local database, were evaluated for different mining scenarios. The mining tasks used the automatic feature identification algorithm which largely aspects the performance of the system. The reason is that with smaller thresholds more features are identified and hence more sentences are analyzed.

VI. CONCLUSION

We used a pattern mining algorithm called H-mine to discover features of products from reviews. It is able to deal with two major issues: 1) Taking many scans of large databases to generate frequent itemsets, and 2) Lack of recognizing transposition of the words while generating new itemsets. In this work we only focused on those features that frequently appear in the review sentences. Our experimental results indicate that our method outperforms the old pattern mining technique used on both precision and recall. Many of the works, including this one, have mainly contributed to discover patterns in the language which can be reused in different cases without binding it to a large number of annotated terms. Opinion mining is a relatively new area of study and very challenging, since it deals primarily with a rather complex system: the human language.

REFERENCES

[1] A.-M. Popes and O. Etzioni, "Extracting Product Features and Opinions from Reviews," in *Proceedings of Human Language Technology*

Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, 2000.

[2] C. Agrawal, R. Agrawal and V. Prasad, "Depth first generation of long patterns," in *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, 2005.

[3] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng and J. Chun , "Red Opal: Product-Feature Scoring from Reviews," in *Proceedings of the 8th ACM conference on Electronic commerce*, New York, 2007.

[4] G. Somprasertsri and P. Lalitrojwong, "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization," *Journal of Universal Computer Science, vol. 16,* pp. 938- 955, 2010.

[5] W. Young Kim, J. Suk Ryu, K. I. Kim and U. Mo Kim, "A Method for Opinion Mining of Product Reviews using Association Rules," in Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, 2009.

[6] M. Hue and B. Liu, "Mining Opinion Features in Customer Reviews," *American Association for Artificial Intelligence,* pp. 755-760, 2012.

[7] R. Hemalatha, A. Krishnan and R. Hemamathi, "Mining Frequent Item Sets More Efficiently Using ITL Mining," in *3rd International CALIBER*, Ahmedabad, 2005.

[8] W. Chin-Ping, C. Yen-Ming, Y. Chin-Sheng and Y. Christopher C., "Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews," in Springer, 2009.

[9] W. Young Kim, J. Suk Ryu, K. I. Kim and U. Mo Kim, "A Method for Opinion Mining of Product Reviews using Association Rules," in Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, 2014