I J C T A, 8(5), 2015, pp. 1973-1980

Enhanced Frequency-based Method for Product Feature Extraction

A. Soundariya*, N. Balaganesh and K. Muneeswaran*****

Abstract: On line reviews play an vital role in online advertising and marketing. Review mining has been given high consideration nowadays, which aims to extract valuable information from the available huge product reviews. Product feature extraction is one of the basic tasks of product review mining. This work is aimed to extract the features from online consumer reviews by considering the nouns and noun phrases that occur frequently which is followed by pruning method to improve the relevance of the extraction result. Finally, the similarity between the feature and the product is identified by using the global search engine thereby enhancing the performance. Experiment is carried out by using the online consumer reviews of electronic products and this procedure can also be used for diverse products domains and datasets.

Index Terms: Feature extraction, Online reviews, Product review mining, PMI-IR.

1. INTRODUCTION

In this new information age, thoughts and opinions about the products are shared publicly through online websites. In order to make best use of this information, we should be able to distinguish what is important and interesting. There are apparent gain to companies, consumers, governments and many organizations in understanding what the public think about their products and services. First, the explosive growth of online product reviews makes it time-consuming and ineffective for consumers to navigate individual reviews one by one in order to obtain useful information about a product, causing an information surplus problem. Second, in reality, individual consumers may have their own preferences for different product features. For example, some consumers may consider appearance and weight as the most significant factors while purchasing a mobile phone, while others may be mainly concerned about its battery life or functions. As a result, it is practically impossible for a consumer to wade through hundreds or even thousands of reviews in order to identify reviews that have commented on specific product features.

Product feature extraction is essential to address the above problem. Feature extraction aims to extract fine-grained features of a product from opinion texts. A feature is a quality of an item, the item may either be a product or service. Feature extraction aims to extract fine-grained excellence of a product from opinion texts. Features are important in sentimental analysis (also called opinion mining), because without knowing them, the opinions expressed on the texts are of limited use. the performance of opinion word detection and sentiment orientation identification.

Likewise feature detection is critical to sentiment analysis, because its effectiveness dramatically affects the performance of opinion word detection and sentiment orientation identification. This work focuses on the extraction of the relevant features. Specifically, given a set of user reviews about a specific product we

* Research Scholar, Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, India, *Email: a.soundariya@gmail.com*

** Research Scholar, Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, India, *Email: balaganesh@mepcoeng.ac.in*

*** Senior Professor, Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, India, *Email: kmuni@mepcoeng.ac.in*

deal with the problem of identifying features on which customers have expressed their opinions. There are various methods available for the extraction of features from the online consumer reviews. The most commonly used approaches are lexicon based feature extraction, dependency relation based feature extraction, and machine language approaches for feature extraction.

In this work, the frequently commented features are considered to be important features. Generally noun and noun phrases are considered as features and are collected from the online consumer reviews. Then, to find out the frequent nouns and noun phrases, an association rule mining technique like apriori algorithm is used which is followed by redundancy pruning for the removal of the single word candidate features. The next step is to remove the proper, personal, brand and verbal nouns and items that are all present along with the features. Finally the Pointwise Mutual Information (PMI-IR) technique is used to identify the semantic similarity between the product features and the item [1]. The main advantage of using PMI-IR is that it searches the web for finding the similarity measures. Since the usage of the online shopping is growing in an exponential manner, the proposed work finds its importance to partially decide the quality of the product being purchased.

The rest of the paper is organized as follows. In section 2 previous works related to our proposed system are discussed, section 3 explains the model and its functionalities, section 4 discuss the results obtained from our work and finally section V conclude the paper.

2. RELATED WORK

The product features are generally extracted from online consumer reviews by number of methods. Many research works have been reported in the literature for extraction of the features and opinions from reviews.

Kovalamudi [2] extracted the features without making use of any of the natural language tools; rather it treats the text as Bag of words and uses the knowledge of Wikipedia. And to do this, they compute the relation from the links to these articles in Wikipedia. Thus for every Wikipedia word, they find the semantic relatedness to all other such words. The main advantage of using Wikipedia is that data freshness is achieved i.e updated data is used. Yahui Xi [3] uses Double Propagation to extract the product feature from Chinese product reviews and adopted some procedures to enhance the precision and recall. Firstly, indirect relations and verb product features are presented to intensify the recall. Secondly, when ranking candidate product features using HITS, he extended the number of hubs by means of the dependency relation patterns among product features and opinion words to improve the precision. Finally, the Normalized Pattern Relevance is used to filter the extracted product features.

Liu and Lim [4] use double propagation method to extract the features of a product from the review. It is an unsupervised method and it works well for medium-sized corpora. However, it is not suitable for small and large sized corpora. To deal with this problem, they introduced two improvement based on part-whole and the well-known web page ranking algorithm HITS is used to find important features. This method achieved improved recall. Ana maria [5] proposes a technique called OPINE an unsupervised information extraction system that extracts opinion phrases, which are adjective, noun, verb or adverb phrases expressing customer opinions. Opine shows improved precision and recall. However, it extracts only explicit features. Hu and Liu [6] uses part whole relation to extract the features and it provides feature-based summary of a large number of customer reviews of a product sold online. It helps in determining infrequent features but fails to find implicit feature.

Hu and Liu [7] laid a method of using Apriori algorithm to mine the features from frequent noun and noun phrases by considering them as product candidate feature. To increase the precision and recall apriori step is followed by the pruning step. Popescu, Nguyen, and Etzioni [8] employed PMI-IR to estimate the semantic associations between feature candidates and discriminator phrases (i.e., part-of and is-a relations)

which in turn is based on the KnowItAll, a web-based, domain-independent information extraction system by Etzioni, Cafarella, Downey, et al [10]. Liu [9] uses the relationship between the feature and the opinion as opinions expressed on some features. Such dependency relationship was first used to find the nearest nouns and noun phrases of a set of well-known opinion words. As a result, this technique is capable of identifying infrequent features missed by the Apriori algorithm. The PMI scores of each noun or noun phrase were then converted into a binary value using a Naive Bayes classifier.

Features can be extracted through supervised learning method [11] in which certain training samples are needed, then supervised learning technique are applied to make an extraction model, which is capable of identifying product features from new customer reviews. Different approaches such as Maximum Entropy, Hidden Markov Models, Conditional Random Fields, Class Association policy and Naive Bays Classifier and extra Machine Learning approaches have been employing for this task. Poria et al. [12] proposed a rule-based approach that exploited common-sense knowledge and sentence dependency trees to identify product features. A predefined implicit feature lexicon was used to extract product features and an opinion lexicon was used to examine users feature tendency. A lexicon-based approach can be easily implemented. However, the effectiveness of this approach highly depends on predefined feature lexicons. With the rapid increase of online products, it is impractical to manually develop and update feature lexicons for every product.

In this work, we propose a technique that extracts features based on frequent nouns and noun phrases and it uses PMI-IR technique to extract the product feature thereby enhancing the performance.

3. PROPOSED SYSTEM

This section provides the detailed description of the proposed system which includes the following steps: PoS tagging, Noun and Noun phrase identification, Mining, Superset based Pruning, Removal of names and items and identification of semantic similarity between the product feature and the product item. The system design is shown in Fig.1.

3.1. Dataset

The dataset is collected from the online review website amazon.in[13]. All the reviews are from the domain of electronic products. The dataset format contains reviewer name, summary of the review, review date, verified purchaser or not, how many consumers are helped by using the review and ratings. The dataset includes three electronic products tablet, television, mp3 player. For each product three brands are collected.

3.2. Part of speech tagging

Generally, the features are the noun and noun phrases with some exceptions. Initially all noun and noun phrases in online consumer reviews are considered as features. The reviews are first split into sentences, and the sentences are parsed using the Part of Speech (POS) tagged tool. POS is a software package that reads text and assigns parts of speech tags to each word, such as noun, verb, adjective, etc. This POS tagging method has a two levels tagger, and the differences are that, level one marks only the noun and verb; whereas level two marks more complicated parts such as adjectives or verbs with noun words function, proper nouns, and morphemes.

In order to increase precision, we used the level two. After POS tagging, basic noun phrases are extracted in accordance with basic noun phrase pattern. Some pre-processing steps which include removal of stop words, punctuation, and numbers are performed.

The used notations are described below.

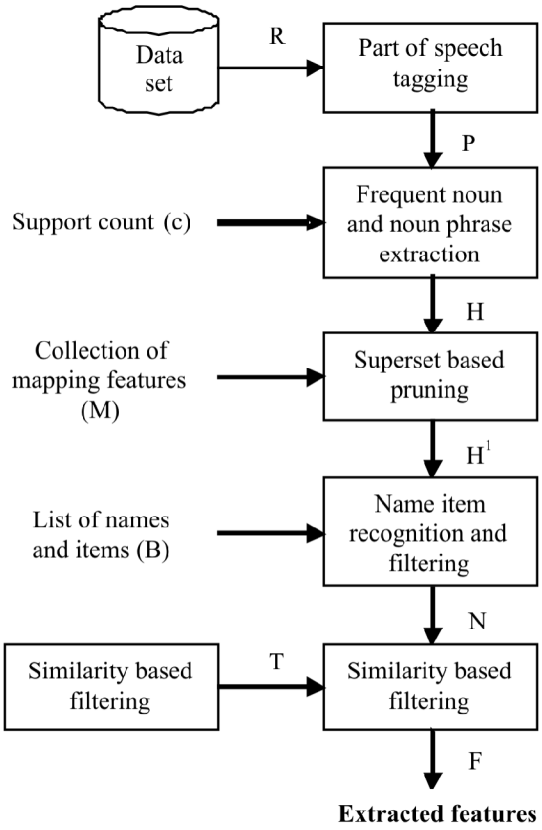

Figure 1: Overall system design

P – Extracted noun and noun phrases
H – Frequent noun and noun phrases
H^1 – Pruned features
N – NIR filtered features
T – Threshold value

3.3. Frequent noun and noun phrase extraction

The frequent features are those features that are reviewed by many customers. For mining frequent items, association rule mining is performed. The reason for using association mining is that a customer review contains lots of words that are not directly related to product features. Different customers usually have different opinions. However, when they mention on product features, the words that they use will converge. Thus using association mining to extract frequent features is essential because those frequent features are probably the product features.

We apply apriori algorithm on the candidate features obtained on the previous step and the resultant features are considered to be the possible relevant features of the products. The Apriori algorithm works in two steps. In the first step, it finds all frequent items from a noun and noun phrase list that satisfy a user-specified minimum support. In the second step, it generates rules from the identified frequent items. For our task, we only need the first step, i.e., finding frequent items, which are candidate features. The support count values changes based on the size of the document.

3.4. Superset based pruning

Superset based pruning is done in order to eliminate the single word candidate features i.e some features may be the subset of another feature. It prunes the single-word candidate features that seldom appear alone,

but often along with other candidate features. One noun can be identified with repeated nouns which may be the subset. So a mapping database is created to obtain the pruned feature. It focuses on removing redundant features that contain single word by mapping with database. We filter single word features which are the superset of the multiple words. For example consider *battery* and *battery life*, the words *life* and *battery* are the subset of the word *battery life* and hence they are considered as single word candidate feature and are pruned.

3.5. Name Item Recognition Filtering

There are some frequent nouns and noun phrases that are used to refer any type of items such as people, place, thing, and event. Those noun and noun phrases are not real features and to filter such kind of words Name Item Recognition (NIR) technique is used. The nouns and noun phrases that belong to the category shown in Table 1 are filtered from the feature list. Name Item recognition consists of two steps detection of names and classification of the names by the type of the item they refer to (e.g. person, organization, location and other).

Table 1
Example of names and items

Types	*Example*
Proper nouns	Paris, January
Brand names	Canon, Dell
Personal nouns	Father, Mother
Verbal nouns	Something, Nothing

3.6. Semantic Similarity based Filtering

$$PMIIR(item,\ feature) = \log_2 \frac{ns(item \wedge\ feature)}{ns(item)ns(\ feature)} \tag{1}$$

To calculate semantic similarity generally wordNet and text corpora are used. However, semantic similarity between words change over time as new senses and associations of words are constantly created. If the technique of wordnet or text corpora is used, the updated semantic similarity cannot be determined and the similarity value will not be appreciable if the size of the corpora is very small. The PMI-IR technique is used to find the semantic association between the product features and the product item using web. PMI-IR score is computed by using the number of searches returned by a search engine and is given in eqn (1).

3.7. Threshold Learning

In order to find the threshold value, the training samples are used. To decide the training samples, from each product domain a brand is taken and the average PMI-IR value of these training samples are considered as the threshold value. The features above the threshold are considered as the actual feature. In this work, three product domains are considered and so from each product domain, a brand is chosen. So the threshold value will be the average of three products. The overall algorithm for the proposed work is shown in fig 2.

Figure 2 describes about the overall process of the proposed system. First the noun and noun phrases (i) are extracted from the online consumer reviews R. Then to extract the frequent noun and noun phrases apriori algorithm (ii) is applied. There may be some redundant features that are all the superset of another features, to remove them pruning (iii) is done by using the mapping features. Names, items location are not tending to be the features. So they are removed by removenameitems (iv). Then the similarity score is calculated between the feature and the item by using (v). To calculate the threshold value average value of

Algorithm RetrieveFeatures (R, c, M, B, F)

Inputs

R-a set of Reviews

c-Support count

M-collection of mapping features

B-List of names and items

Output

F-Retrieved Features

i) P = extractNNP (R)

ii) H = applyApriori (P, c)

iii) H^1 = prune (H, M)

iv) N = removenameitems (H^1, B)

v) PMIR = computePMI_IR (N) // by equation (1)

vi) T = average (PMIR)

vii) F = PMIR > T

PMIR-PMI-IR score for extracted features

Figure 2: Overall Algorithm

PMIR is considered (vi). The features above the threshold are considered as actual product features as provided in (vii).

4. RESULTS AND DISCUSSIONS

This section discusses about the intermediate results of the each component of the system and provides the performance evaluation of the implemented methodologies.

4.1. Experimental Setup

Table 2 shows the list of products along with their domain that were used for our experiment. In this work three elecronic products of different brands are considered.

Table 2
List of product Details

Tablet	*Television*	*Mp3 Player*
Samsung-Galaxy	Samsung-23H4003	Zebronics-Node
iBall-Performance	LG-32LB550A	Transcend-MP350
Dell-Venue	Panasonic-29PFL4738	Philips-Go Gear

The average of PMI-IR score 29.86 is used as a threshold value and the features with the score above this threshold is considered as the actual features.

4.2. Performance Analysis

In order to evaluate the performance of extracted features Precision, Recall and F-Score are used. Precision is the fraction of selected product features that are correct and recall is defined as the fraction of correct product features that are selected. F-score is the harmonic average of precision and recall. It gives the percentage of correct features that are among the top *N* feature candidates in an extracted list and are given in eqn (2), (3) and (4). The lists of relevant features are obtained manually by identifying the features from the customers product review.

Table 3 shows the performance analysis for the product reviews. It shows the precision, recall and f-score of all the products.

$$precision = \frac{retrived \quad features \cap relevant \quad features}{retrived \quad features} \tag{2}$$

$$recall = \frac{retrived \quad features \cap relevant \quad features}{relevant \quad features} \tag{3}$$

$$F - score = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

Table 3
Performance Analysis

Product Names	*PMI-IR*			*PMI*		
	Precision	*Recall*	*F Score*	*Precision*	*Recall*	*F Score*
Samsung tab	71	75	73	72	73	73
iBall	75	67	71	72	69	70
Dell-Venue	71	67	69	67	67	67
Samsung TV	68	67	67	68	69	69
LG	78	64	70	77	62	69
Panasonic	72	81	76	74	75	74
Zebronics	71	71	71	70	69	69
Transcend	82	64	72	79	63	70
Philips	72	71	71	72	67	69

The method of PMI-IR is compared with the PMI technique. The method of PMI-IR checks the similarity based on the search engine searches. But PMI searches the similarity only among the review sentences. The performance analysis implies that PMI-IR works well when comparing with that of PMI method. It can be inferred from the Table 3 that the precision is good for the PMI-IR technique rather than the PMI technique. As there is a good improvement in the precision, the retrieved features are more likely to be the relevant features. The value of recall implies that the relevant features are retrieved. Overall, the PMI-IR method works better than the PMI method.

5. CONCLUSION AND FUTURE WORK

The method proposed in this work extracts the frequent features available in the review by calculating semantic similarity between features and item, for which the technique of PMI-IR is used. Advantage of using

PMI-IR technique is that it uses the global search engine to calculate semantic similarity. The extracted feature can be used by the manufacturers to cluster the user review based on product feature and obtain positive and negative opinion on the product features. It can also be used by the consumers to make better purchase decision and to compare different product based on the product feature. In this work only frequent features are considered to be the feature. To further increase the performance, infrequent features and implicit features should be considered and refinement should be made on used techniques**.**

Acknowledgement

The authors wish to thank the Management and Principal of Mepco Schlenk Engineering College, for the support in carrying out this research work.

References

[1] Shi Li, Lina Zhou, Yijun Li, "Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures," Elsevier Transaction on Information Processing & Management, Volume 51, Issue 1, January 2015, Pages 58-67.

[2] Kovelamudi, S., Ramalingam, S., Sood, A., & Varma, V., "Domain independent model for product attribute extraction from user reviews using Wikipedia," in Proceedings of the 5th international joint conference on natural language processing, Chiang Mai, Thailand, August 2011, pp. 1408-1412.

[3] Yu Zheng; Liang Ye; Geng-feng Wu; Xin Li, "Extracting product features from chinese customer reviews," Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on, Nov. 2008 vol. 1, no., pp. 285,290, 17-19.

[4] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain, "Extracting and ranking product features in opinion documents," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, Stroudsburg, PA, USA, August 2010, 1462-1470.

[5] Ana-Maria Popescu and Oren Etzioni, "Extracting product features and opinions from reviews," in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2005, 339-346.

[6] Minqing Hu and Bing Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2004, 168-177.

[7] Minqing Hu and Bing Liu, "Mining opinion features in customer reviews," in *Proceedings of the 19th national conference on Artifical intelligence*, Anthony G. Cohn (Ed.), AAAI Press, 2004, 755-760.

[8] Popescu, A. M., Nguyen, B., & Etzioni, O., "OPINE: Extracting product features and opinions from reviews," in HLT-Demo '05: Proceedings of HLT/EMNLP on interactive demonstrations, Association for Computational Linguistics, 2005, 32-33.

[9] Liu, B, "Feature-based sentiment analysis in Sentiment analysis and opinion mining," Morgan & Claypool Publishers, 2012, pp. 16-28.

[10] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates, "Unsupervised named-item extraction from the web: an experimental study," *Artif. Intell.* 165, 1, June 2005, 91-134.

[11] S. Revathi Manju, E. V. R. M. Kalaimani, R. Bhavani, "Product Feature Ranking Using Semantic Oriented Sentiment Classifier", in International Journal of Scientific Engineering and Research, Volume 2 Issue 10, October 2014.

[12] S. Poria, E. Cambria, L.-W. Ku, C. Gui, A. Gelbukh, "A Rule-based Approach to Aspect Extraction from Product Reviews", in the Second Workshop on Natural Language Processing for Social Media, Dublin, Ireland, 2014 pp. 28–37.

[13] The Dataset adopted for this work available on [Online]: http://amazon.in