Product Aspect Identification and Ranking System

Rutuja Tikait, Prof. Ranjana Badre, Prof. Mayura Kinikar

Abstract*— Online retail is growing by leaps and bounds. Many retail websites and forum websites contain thousands of customer reviews about various product expressing their opinion on various aspects of product . These reviews are rich in knowledge but are disorganize creating problem in information navigation and knowledge acquisition . to address this problem this paper proposes a system to automatically identify important aspects of products from online reviews . aspect ranking algorithm is proposed using the concept TFIDF associated with occurrence probability of opinionated words for calculation of weight of aspects. This algorithm is experimented using product review dataset to show the effectiveness of the ranking approach. The scope of the proposed system is to organize the consumer review in appropriate way so that the product promotion can be done effectively based on reviews.

***Index Terms*— aspects; aspect ranking; aspect identification; consumers; consumer review; sentiment classification.**

I. Introduction

Use of internet and e-commerce is growing very rapidly. Many products are available online. Most of the retail Websites encourages consumers to write their feedbacks about products to express their opinions on various *aspects* of the products. This gives rise to huge collection of reviews on web. These reviews contain rich and valuable knowledge and have become an important resource for both consumers and firms. Consumers commonly look for quality information from online reviews before buying a product and firms can use these reviews as feedback for better product development, consumer relationship management and for the development of new marketing strategies.

A product may have hundreds of aspects. Some of the product aspects are more important than the others and have strong influence on the eventual consumer’s decision making as well as firm’s product development strategies. Identification of important product aspects become necessary as both consumers and firms are benefited by this. Consumers can easily make purchasing decision by paying attention to the important aspects as well as firms can focus on improving the quality of these aspects and thus enhance product reputation efficiently.

Simple solution for important aspect identification is frequency based which selects frequently commented aspects in consumer’s reviews as important. However, consumers’ opinions on the frequent aspects may not influence their overall opinions on the product, and thus not influence consumers’ purchase decisions.

Motivated by the above observations, this paper presents a product aspect ranking system which can identify important product aspects from consumer reviews and rank them by taking into account the frequency and consumer’s opinion on frequent aspects. Aspects will be ranked by using weight calculated for the aspects. TFIDF algorithm is commonly used for weight calculation. TFIDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. In this paper a new aspect ranking algorithm is used which makes the use of TFIDF to calculate the value of aspect and also calculate the occurrence frequency of opinionated words associated with aspect for the calculation of weight. Aspects will be ranked by using weight calculated by using this proposed method.

II. Related work

This section briefly survey previous work on product aspect ranking framework starting with the product aspect identification. Existing product aspect identification techniques can be broadly classified into two main approaches: supervised and unsupervised [2].

Supervised learning technique learns an extraction model which is called as aspect extractor, that aspect extractor is then used to identify aspects in new reviews. For this task Hidden Markov Models and Conditional Random Fields [3, 4], Maximum Entropy [13], Class Association Rules and Naive Bayes Classifier [14] approaches have been used. Wong and Lam [3] used a supervised learning technique to train an aspect extractor. They learned aspect extractor using Hidden Markow Model and conditional random field. Supervised techniques is reasonably effective, but preparation of training examples is time consuming.

In contrast, unsupervised approaches automatically extract product aspects from customer reviews without using training examples. Hu and Liu's works [5, 6] focuses on association rule mining based on the Apriori algorithm to mine frequent itemsets as explicit product aspects. In association rule mining, the algorithm does not consider the position of the words in the sentence. In order to remove incorrect frequent aspects, two types of pruning criteria were used: compactness and redundancy pruning. The technique is

***Manuscript received April, 2014*.**

Rutuja Tikait, *Department of Computer Engineering, MIT Academy of Engineering., Pune, India, Mobile No:9763967282.*

Prof. Ranjana. Badre*, Department of Computer Engineering, MIT Academy of Engineering., Pune, India,Mobile No:9822247585.,*

Prof. Mayura Kinikar*, Department of Computer Engineering, MIT Academy of Engineering., Pune, India, Mobile No:9823391166.,*

efficient which does not require the use of training examples or predefined sets of domain-independent extraction patterns. However, it suffers from two main shortcomings. First, frequent aspects discovered by the mining algorithm might not be product aspects. The compactness and redundancy pruning rules are not able to eliminate these false aspects. Second, even if a frequent aspect is a product aspect, customers may not be expressing any subjective opinion about it in their reviews.

Wu et al [7] also used the unsupervised method. They used the phrase dependency parser to extract noun and noun phrases and then they used a language model to filter out the unwanted aspects. This language model was used to predict the related score of candidate aspects and was built on product reviews. Candidate having low score were filtered out. However this language model might be biased to frequent terms in the reviews and cannot predict the aspect score exactly as a result cannot filter out noise very efficiently. Subsequently, Popescu and Etzioni [17]developed the *OPINE* system, which extracts aspects based on the *KnowItAll*Web information extraction system [18].

After identification the important aspects next step is sentiment classification which is used to determine the orientation of sentiment on each aspects. Aspect sentiment classification can be done by using two approaches first Lexicon based approach and second supervise learning approach. Lexicon based approach is typically unsupervised. Lexicon consists of list of sentiment words, which may be positive or negative. This method usually employs a bootstrap strategy to generate high quality Lexicon. Hu and Liu [5] have used this lexicon based method. They obtained the sentimental lexicon by using synonym/antonym relation defined in WordNet to bootstrap the seed word set.

Hu's method is improved by Ding et al [8] by addressing two issues: opinion of sentiment word would be content sensitive and conflict in review. They derived the lexicon by using some constraints.

Second approach is supervised learning approach which classifies opinions on aspects by using sentiment classifier. Sentiment classifier is learned from training corpus which is used to classify the new aspects opinions. Many learning models are applicable for this purpose [9]. Bopong and Lee [10] used 3 machine learning techniques SVM, Naïve Bayes and Maximum Entropy for determining whether the review is positive or negative.

The product aspect ranking is to predict the ratings on individual aspects. Wang *et al*. [15] developed a latent aspect rating analysis model, which aims to infer reviewer's latent opinions on each aspect and the relative emphasis on different aspects. This work concentrates on aspect-level opinion estimation and reviewer rating behavior analysis, rather than on aspect ranking. Snyder and Barzilay [16] formulated a multiple aspect ranking problem. Justin Martineau and Tim Finin [19] present Delta TFIDF, a general purpose technique to efficiently weight word scores. This technique calculate the value of aspect in document but does not take into account the frequency of words associated with aspect with it.

III. System architecture

Proposed product aspect ranking system takes the customer reviews dataset as input and perform various steps on the dataset to generate output. System performs preprocessing on dataset then aspect identification, sentiment classification and finally apply aspect ranking algorithm to provide ranking of aspects. Fig. 1 shows the architecture of system.

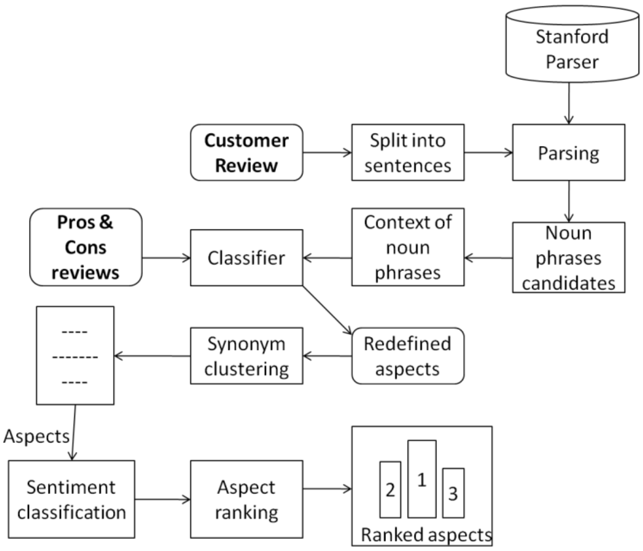

Fig. 1 Product aspect ranking System Architecture.

A. *List of Modules*

- Product Aspect Identification
- Aspect-level Sentiment Classification
- Probabilistic Aspect Ranking
- Evaluation

a) Module 1 (Product Aspect Identification)
Input:
Pros and cons reviews and
Free text reviews (Amazon and Flipkart)
Output:
Product aspects.

In the Pros and Cons reviews, the aspects are identified by extracting the frequent noun terms in the reviews. For identifying aspects in the free text reviews, first the free text reviews are spilt into sentences, each sentence is parsed using Stanford parser. The frequent noun phrases are then extracted, with the help of above mentioned function, as candidate aspects.

Each aspect in the Pros and Cons reviews are represented into a unigram feature, and utilize all the aspects to learn a one-class Naive Bayes classifier. Stanford parser gives a parse tree as its output, from which noun phases has to be extracted.

Product aspects can comprise of only nouns and adjectives.

```
function getRequiredWords(node)
if node==leaf_node and (POS_node == NN or POS_node == JJ)
    add the node to the result list
else
    for each child of node
        getRequiredWords(child)
```

Using this classifier, product aspects are identified. Context analysis is also done for better classification. As the identified aspects may contain some synonym terms, synonym clustering is done to obtain unique aspects. The synonym terms are collected from the synonym dictionary Website.

b) Module 2 (Sentiment Classification on Product Aspects)

Input:
Collection of reviews and identified aspects.
Output:
The customer's opinion on specific aspects is found for each aspect.

A Sentiment classifier is learned from the Pros reviews (positive reviews) and cons reviews (negative reviews). The classification is done using Naive Bayes model classifier. The Pros and Cons reviews have explicitly categorized positive and negative opinions on the aspects. These reviews are valuable training samples for learning a sentiment classifier.

Pros and Cons reviews are used to train a sentiment classifier, which is in turn used to determine consumer opinions on the aspects in free text reviews. First sentiment terms in Pros and Cons reviews are collected, then the classifier is trained using these sentiment terms and this trained classifier is used to classify the aspect in free text review.

c) Module 3 (Product Aspect Ranking)

Input:
Collection of reviews and identified aspects.
Output:
The customer's opinion on specific aspects is found for each aspect.

Proposed aspect ranking algorithm calculates the weight of aspects of a product from consumer reviews. This algorithm uses the concept of TFIDF which is commonly used for calculation of weight of term in document. Here this concept is used for calculation of value of aspect term. Weight of aspect is calculated by using aspect value given by TFIDF and occurrence frequency of positively opinionated word and negatively opinionated words associated with aspect term.

B. Proposed Aspect Ranking Algorithm

a) Terms used in Algorithm

- D={r1, r2,r3…rn} be the set of reviews.
- A_k= {a1, a2, a3,…… an} be the set of aspect
- $C_{a,D}$ is the number of times aspect term a occurs in review dataset D.
- P_a is the number of comments in the positively labeled set with aspect term a.
- |P| is the number of comments in the positively labeled set.
- N_a is the number of comments in the negatively labeled set with aspect term a.
- |N| is the number of comments in the negatively labeled set.
- $V_{a,D}$ is the feature value for aspect term a in review dataset D.
- Let Φ = set of positive words
  Φ= {P1, P2, P3 ... Pn}
- Let ψ = set of negative words
  ψ ={N1, N2, N3 ... Nn}
- $P(\Phi)$=probability of Φ
- $P(\psi)$=probability of ψ
- ω weight of aspect a

b) Algorithm Steps

- Calculate *the value of aspect a, given by*

$$\begin{aligned} V_{a,D} &= C_{a,D}*\log_2(|P|/ P_a) - C_{a,D}*\log2(|N|/ N_a) \\ &= C_{a,D}*\log_2((|P|\, N_a /P_a|N|) \\ &= C_{a,D}*\log_2(N_a / P_a) \end{aligned}$$

- Calculate *the occurrence probability of each positively opinionated word.*

$$\alpha = \sum_{i=1}^{n}(\mathrm{P}(\Phi i) * \mathrm{W}(\Phi i))$$

- Calculate *the occurrence probability of each negatively opinionate word.*

$$\beta = \sum_{i=1}^{n}(\mathrm{P}(\psi i) * \mathrm{W}(\psi i))$$

- Calculate *weight,*

$$\omega = \mathrm{V}_{a,D} - \sum_{i=1}^{D}(\alpha - \beta)$$

IV. DATA REQUIREMENTS

The only data required for this project are the product reviews of customers. Reviews can be posted on the webs in three different types :

Type(1)- Pros and Cons: The reviewer is asked to describe Pros and Cons separately.

Type (2)- Pros, Cons and detailed review: The reviewer is asked to describe Pros and Cons separately and also write a detailed review.

Type (3)- Free format: The reviewer can write freely, i.e., no separation of Pros and Cons. Different types of reviews may need different techniques to perform the tasks such as product aspect identification, product sentiment classification and product aspect ranking.

For This project the reviews are taken From from amazon.com for six product and the details of dataset are as follows:

Sr. no.	Product Name	No. of reviews
1.	Mobile	450
2.	Camera	400
3.	Laptop	350
4.	Ipod	300
5.	Tv	375
6.	Printer	300

Table. 1 Details of input Dataset

V. Evaluation of Aspect Ranking

To evaluate the performance of aspect ranking, we adopted the widely used *Discounted Cumulative Gain* at top *k* (DCG@*k*) as the evaluation metric. Given a ranking list of aspects, DCG@*k* is calculated as

$$DCG@k = \sum_{i=1}^{k} \frac{2^{t(i)} - 1}{\log(i + 1)}$$

where *t(i)* is the importance degree of the aspect at position *i*, and *Z* is a normalization term derived from the top-*k* aspects of a perfect ranking. For each aspect, its importance degree was judged by three annotators as three importance levels, i.e. "*Un-important*" (score 1), "*Ordinary*" (score 2), and "*Important*" (score 3).

In order to evaluate the effectiveness on aspect ranking, proposed aspect ranking algorithm is compared with the following three methods:

(a) Frequency-based method:
It ranks the aspects according to aspect frequency.

(b) TFIDF Based method:
This technique calculate the value of aspect in document but does not take into account the frequency of words associated with aspect with it.

Fig. 2 Shows the result of comparison of above two methods with proposed method.

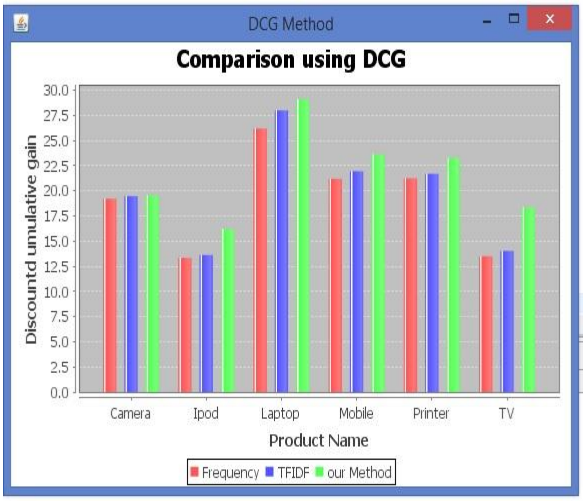

Fig.2 Performance of aspect ranking in terms of DCG

On average, the proposed aspect ranking approach significantly outperforms frequency-base and TFIDF based method in terms of DCG by over 9.7% and 6.8% respectively. Hence, we can speculate that the proposed approach can effectively identify the important aspects from consumer reviews by simultaneously exploiting aspect frequency and the influence of consumers' opinions given to each aspect over their overall opinions. The frequency-based method only captures the aspect frequency information, and neglects to consider the impact of opinions on the specific aspects on the overall ratings. It may recognize some general aspects as important ones. Although the general aspects frequently appear in consumer reviews, they do not greatly influence consumers' overall satisfaction.

Where as in TFIDF based method calculate the value of aspect in document by considering opinion as either positive or negative but does not take into account the frequency of opinionated words associated with aspect with it.

VI. Conclusion

A product aspect ranking System is used to identify the important aspects of products from numerous consumer reviews. System contains three main components, i.e. product aspect identification, aspect sentiment classification, and aspect ranking. First, system used the Pros and Cons reviews to improve aspect identification and sentiment classification on free-text reviews. Then an aspect ranking algorithm is used to calculate the weight of various aspects of a product from numerous reviews. The product aspects are finally ranked according to their weight. Proposed method shows the performance improvement over the two existing systems frequency based system and TFIDF based system in terms of DCG by 9.7% and 6.8% respectively.

Acknowledgment

I express true sense of gratitude towards my project guide **Prof. Ranjana Badre and Prof. Mayura Kinikar**, of computer department for her invaluable co-operation and guidance that she gave me throughout my research. I specially thank our **P.G coordinator Prof. R. M. Gaudar** for inspiring me and providing me all the lab facilities, which made this research work very convenient and easy. I would also like to express my appreciation and thanks to our HOD **Prof.Uma Nagaraj** and Principal **Dr. Y. J. Bhalerao** and all my friends who knowingly or unknowingly have assisted me throughout my hard work.

References

[1]. Zheng-JunZha, Jianxing Yu, Jinhui Tang,Meng Wang, and Tat-Seng Chua, "Product AspectRanking and Its Applications", IEEETRANSACTION ON KNOWLEDGE AND DATA ENGINEERING, vol.26,no.5, May 2014

[2]. Rutuja Tikait, Ranjana Badre and Mayura Kinikar, "Product aspect Ranking Techniques A Survey", IJIRCCE, Nov 2014.

[3]. T. L. Wong and W. Lam, "Hot item mining and summarization from multiple auction web sites," in *Proc. 5th IEEE* ICDM,Washington, DC, USA, 2005, pp. 797–800

[4]. Wong, T.L., Lam, W.: Learning to extract and summarize hot item features from multiple auction web sites. Knowl. Inf. Syst. 14(2), lexical and syntactic features. In: Proc. of the IEEE International 143{160 (2008)

[5]. M. Hu and B. Liu, "Mining and summarizing customer reviews," inProc. SIGKDD, Seattle, WA, USA, 2004, pp. 168–177.

[6]. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proc. of American Association for Artificial Intelligence Conference. pp. 755{760 (2004).

[7]. Y. Wu, Q. Zhang, X. Huang, and L. Wu, "Phrase dependency parsing for opinion mining," in Proc. ACL, Singapore, 2009, pp. 1533–1541.

[8]. X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in Proc. WSDM, New York, NY, USA, 2008, pp. 231- 240.

[9]. B. Liu, Sentiment Analysis and Opinion Mining. Mogarn& ClaypoolPublishers, San Rafael, CA, USA, 2012.

[10]. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proc.EMNLP, Philadelphia, PA, USA, 2002, pp. 79–86.

[11]. J. Yu, Z.-J. Zha, M. Wang, and T. S. Chua, "Aspect ranking: Identifying important product aspects from online consumer

reviews," in Proc. ACL,Portland, OR, USA, 2011, pp. 1496–1505.

[12]. B. Ohana and B. Tierney, "Sentiment classification of reviewsusing SentiWordNet," in Proc. IT&T Conf., Dublin, Ireland, 2009.

[13]. Somprasertsri, G., Lalitrojwong, P.: Automatic product feature extraction fromonline product reviews using maximum entropy with Conference on Information Reuse and Integration. pp. 250{255. IEEE Systems, Man, and Cybernetics Society (2008).

[14]. Yang, C.C., Wong, Y.C., Wei, C.P.: Classifying web review opinions for consumerproduct analysis. In: Proc. of the 11thInternational Conference on Electronic Commerce. pp. 57{63. ACM, New York, NY, USA (2009)

[15]. H. Wang, Y. Lu, and C. X. Zhai, "Latent aspect rating analysis on review text data: A rating regression approach," in Proc. 16th ACM SIGKDD, San Diego, CA, USA, 2010, pp. 168–176.

[16]. B. Snyder and R. Barzilay, "Multiple aspect ranking using the good grief algorithm," in Proc. HLT-NAACL, New York, NY, USA, 2007, pp. 300–307.

[17]. A. M. Popescu and O. Etzioni, "Extracting product features andopinions from reviews," in Proc. HLT/EMNLP, Vancouver, BC,Canada, 2005, pp. 339–346.

[18]. O. Etzioni*et al*., "Unsupervised named-entity extraction from theweb: An experimental study," *J. Artif. Intell.*, vol. 165, no. 1, pp.91–134. Jun. 2005.

[19]. Justin Martineau, and Tim Finin" Delta TFIDF: An Improved Feature Space for Sentiment Analysis", Proceedings of the Third International ICWSM Conference ,2009.

Rutuja V. Tikait completed B.E. in Information Technology from Amravati University and pursuing her M.E. degree in Computer Engineering in the Department of Computer Engineering in MIT Academy of Engineering, Pune.

Prof. R.R. Badre received her M.E. degree in Computer Science and Engineering from Shivaji University, Kolhapur. She is currently working as an Associate Professor in the Department of Computer Engineering in MIT Academy of Engineering, Pune. She has published more than 8 papers in both International journals and conferences. She is also a member in Indian Society of Technical Education.

Prof. Mayura Kinikar received her M.E. degree Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. She is currently working as an Assistant Professor in the Department of Computer Engineering in MIT Academy of Engineering, Pune.