ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Product Aspect Ranking Techniques: A Survey

Rutuja Tikait, Ranjana Badre, Mayura Kinikar

P.G. Scholar, Dept. of C.S.E., MIT AOE, Savitribai Phule University of Pune, Pune, India.

Associate Professor, Dept. of C.S.E., MIT AOE, Savitribai Phule University of Pune, Pune, India.

Assistant Professor, Dept. of C.S.E., MIT AOE, Savitribai Phule University of Pune, Pune, India.

ABSTRACT: A product may have hundred 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. This paper provides the description of various techniques for product aspect identification and classification.

KEYWORDS: aspect identification; aspect Ranking; Consumer review; Product aspects; Sentiment classification.

I. INTRODUCTION

In the recent years the use of e-commerce is grown very rapidly. Most retail Websites promotes consumers to write their feedbacks about products to express their opinions on various *aspects* of the products. An *aspect*, which can also be called as *feature*, refers to a component or an attribute of a certain product. A sample review "*The sound quality of Sony Experia is amazing.*" reveals positive opinion on the aspect "*sound quality*" of product *Sony Experia*. Many forum Websites also provide a platform for consumers to post reviews on number of products. For example, CNet.com involves more than seven million product reviews; These numerous consumer reviews contain rich and valuable knowledge, which is becoming an important resource for both consumers and firms [1]. Before purchasing a product, consumers commonly seek quality information from online reviews and firms can use these reviews as feedbacks for better product development, consumer relationship management and marketing.

Generally, a product may have number of aspects. For example, a *Smart Phone* has hundreds of aspects such as "screen size," "camera," "memory size," "sound quality." one may say that some aspects are more important than the others, and have strong influence on the consumers' decision making as well as firms' product development strategies. For example, some aspects of *Smart Phone* e.g., "camera" and "memory size," are considered important by most of the consumers, and are more important than the others such as "color" and "*buttons*." Hence, the identification of important product aspects plays an essential role in improving the usability of reviews which is beneficial to both consumers and firms. Consumers can easily make purchasing decision by paying attention to the important aspects, while firms can focus on the improvement of product quality so that product reputation is enhanced. However, manual identification of important aspects is impractical. Therefore, an approach to automatically identify the important aspects is highly demanded. Motivated by the above observations, we made a survey on different techniques used to find important product aspects automatically from online consumer reviews.

In this paper we present the methodology in section II and techniques used for the product aspect identification and product aspect classification in the section no. III and section no. IV respectively and section V illustrates the product aspect ranking.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

II. METHODOLOGY

The process of product aspect ranking consisting of three main Steps: (a) aspect identification; (b) sentiment classification on aspects (c) Product aspect ranking. Given the consumer reviews of a product, first identify the aspects in the reviews and then analyze these reviews to find consumer opinions on the aspects via a sentiment classifier and finally rank the product based on importance of aspect by taking into account aspect frequency and consumers' opinions given to each aspect over their overall opinions.

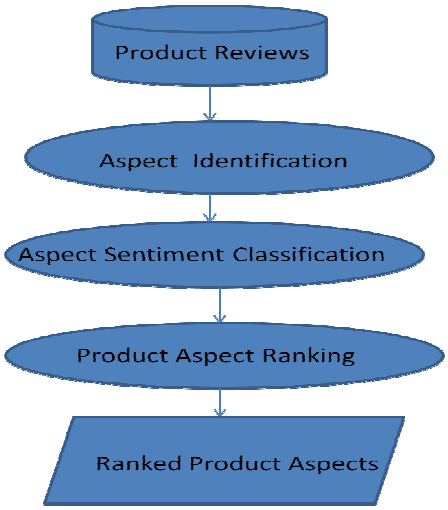

Fig. 1 Product aspect ranking process

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 as mentioned in fig. 1

For type (1) and (2), opinion orientations are known because Pros and Cons are separated and thus there is no need to identify them. Only product features need to be identified from the comments of customers. For type (3), we need to identify both product features and opinion orientations.

In order to obtain identification of aspects, the *Pros* and *Cons* reviews are used as supporting knowledge to assist the identification of aspects in the free text reviews. In particular, first split the free text reviews into sentences, and parse each sentence using parser. After that the frequent noun phrases are extracted from the sentence parsing trees as candidate aspects. Since these candidate aspects may contain noises, further the *Pros* and *Cons* reviews are used to assist them in identification of aspects from the candidates. Then all the frequent noun terms extracted from the *Pros* and *Cons* reviews are collected to form a vocabulary. Each aspect in the *Pros* and *Cons* reviews is represented into a unigram feature, and all the aspects are then used to learn a one-class Support Vector Machine (SVM) classifier [2]. The resultant classifier is used to identify aspects in the candidates extracted from the free text reviews.

This task of analyzing the sentiments expressed on aspects is called aspect-level sentiment classification [3]. Many techniques are used for sentiment classification which includes the supervised learning approaches and unsupervised approaches such as the lexicon-based approaches. The lexicon-based method uses a sentiment lexicon which contains a list of sentiment words, phrases and idioms, to determine the sentiment orientation on each aspect [4]. On the other

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

hand, the supervised learning methods train a sentiment classifier by using training dataset. The classifier is then used to predict the sentiment on each aspect. In the subsequent sub sections we will discuss the various methods for aspect sentiment classification. Finally a probabilistic aspect ranking algorithm is used to identify the important product aspects from reviews.

III. ASPECT IDENTIFICATION TECHNIQUES

A. *Supervised learning:*

Supervised learning technique use the collection of labeled reviews to learn an extraction model. This extraction model called as extractor is then used for the identification of aspects in ne reviews. Most of the supervised learning techniques are based on the sequential learning. Various literatures show the different technique for the learning of extractor.

Wong and Lam [5] uses the HMM model and conditional random field to learn the extractor.

Li et al [18] uses skip CRF and tree CRF i.e. to integrated CRF variation to learn the extractor. The main disadvantage of this method is that it require Labeled sample for training. These methods are very time consuming to label the samples.

B. *Unsupervised learning:*

In this method the aspects are considered noun or noun phases and occurrence frequency of noun and noun phrases is calculated. The frequent noun or noun phrases are considered as aspects. Hu and Liu [3] use this unsupervised technique for aspect identification. Main disadvantage of this method is that identified aspects candidates may contain noise.

Wu et al [21] uses a phrase dependency parsing. Phrase dependency parsing takes the sentence as input and segments it into phrases. Then these segments are linked with directed arc. Phrase dependency parsing focuses on phrases and not on single word inside phrase. To make sure that identified aspects candidate is to be aspect language model which is based on product reviews is used to predict score of candidates. Model filter out low score candidates. Such model may be biased to frequent terms in the review and cannot sense precisely the related score of aspect terms as a result cannot filter out noise efficiently.

Popesu and Etrioni [19] developed their own systems for aspect identification. They developed OPINE system which is based on KnowItAll web information extraction system which extract aspects from reviews. Su et al [20] design a reinforcement strategy. This strategy cluster together product aspect and opinion words by iteratively using both content and sentiments

IV. ASPECT SENTIMENT CLASSIFICATION TECHNIQUES

A. *Lexicon Based Approach :*

Opinion words are used in many sentiment classification tasks. Desired states are express with Positive opinion words while, undesired states are expressed with negative opinion words. All opinion words, opinion phrases and idioms are together called as Lexicon.

Hu and Liu [3] use this method. They utilize synonym /antonym relations defined in WordNet to bootstrap the seed word set and finally obtain a sentiment lexicon. Fig 2 shows example of Bipolar adjective structure where synonym relationship is represented by arrow (→) and antonym relationship is represented by dash arrow (--->).

Three approaches used to collect this opinion word list are manual, dictionary based and corpus based approaches. Manual approach is very time consuming and it is not used alone. It is usually combined with the other two automated approaches to avoid mistakes that resulted from automated methods. The two automated approaches are presented in the following subsections.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

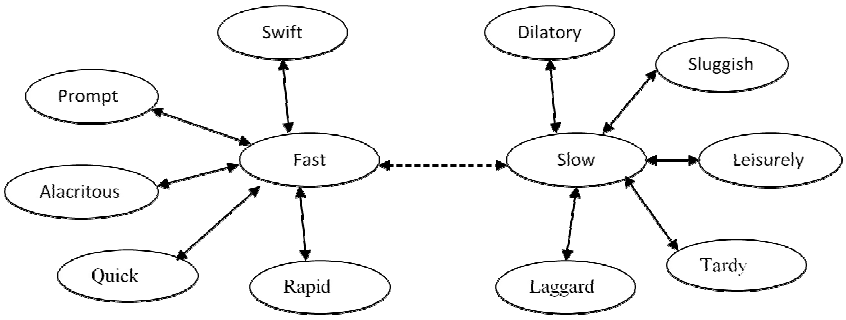

Fig. 2 Bipolar adjective structure,
(→ = similarity; ---> = antonymy)

a) Dictionary-based approach:

[7,8] presented the main strategy of the dictionary-based approach. A small set of opinion words is collected manually with known orientations. Then, synonyms and antonyms of these words is added to this set which is grown by searching the words in the well known corpora WordNet [6] or thesaurus [9]. The newly found words are added to the seed list then the next iteration starts. This iterative process stops when no new words are found. After completion of process list is checked manually to remove or correct errors.

This method is unable to find opinion words with domain and context specific orientation which is the main drawback of this method.

b) Corpus-based approach:

The drawback of dictionary based approach is overcome in Corpus- based approach which helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns.

B. *Holistic Lexicon Based Approach:*

Ding et al [22] presented a Holistic lexicon based method Holistic lexicon-based approach improves the lexicon-based method in [23] by addressing two issues that the opinion of sentiment words would be content sensitive and conflict in the review.

This method does not look at the current sentence alone rather it uses the external information and evidences in other sentence and other reviews. Some linguistic conventions in natural language expression are used to find the orientation of opinion word. This method required prior domain knowledge or user inputs are needed. This approach is highly effective when sentence contain multiple contradictory opinion words.

C. *Supervised Learning Techniques:*

a) Naïve Bayes Classifier (NB):

Naïve Bayesian networks are composed of acyclic graph with only one parent and several children. There is a very strong assumption of independence with child nodes in the context of their parents. Independence model can be represented with :

$$R = \frac{P(i/X)}{P(j/X)} = \frac{P(i)P(X/i)}{P(j)P(X/j)} = \frac{P(i)nP(X/i)}{P(j)nP(X/j)}$$

When these two probabilities are compared, larger probability is more likely to be the actual class label. Advantage of naive bayes classifier is its short computational time for learning the dataset. Bayes classifiers are usually less accurate than that of other learning algorithms.

b) Maximum Entropy Classifier (ME):

Another classifier is Maximum Entropy classifier. The name Maximum Entropy comes from the fact that the classifier finds the probabilistic model which is the simplest and least constrained. Yet it has some specific constraints. The idea behind maximum entropy is that one should prefer the most uniform models that also satisfy any given

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

constraints. This Classifier is used to converts labeled feature sets to vectors using encoding. This encoded vector is then used to compute weights for each feature that can then be combined to determine the most likely label for a feature set. This classifier is controlled by a set of X{weights}, which is used to combine the joint features that are produced from a feature-set by an X{encoding}. Each C{(featureset, label)} pair is mapped to vector with encoding scheme. The probability of each label is calculated using the following equation:

$$P\left(\frac{fs}{label}\right) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l))forlinlabels)}$$

ME classifier was used by Kaufmann [25] to detect parallel sentences between any language pairs with small amount of training data. Other tools were developed to automatically extract parallel data from non-parallel corpora use language specific techniques or require large amounts of training data. Their results showed that ME classifiers can generate useful results for almost any language pair. This can allow the formation of parallel corpora for many new languages.

c) Support Vector Machines Classifiers (SVM):

Support Vector Machines (SVMs) are the newest supervised machine learning technique. SVM uses the notion of a "margin"- a hyperplane that divide two data classes.. An upper bound on the expected generalization error can be reduced by maximizing the margin and thereby largest possible distance between separating hyperplane instances on either side of it. In Fig. 3, X, O are 2 classes and A, B and C are the three hyperplanes. Hyperplane A provides the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation. In the case of linearly separable data, once the optimum separating hyperplane is found, data points that lie on its margin are known as support vector points whose solution is represented as a linear combination of only these points . Other data points are ignored. Therefore, the model complexity of an SVM is unaffected by the number of features encountered in the training data. For this reason, SVMs are well suited for learning tasks where the number of features is large with respect to the number of training instances.

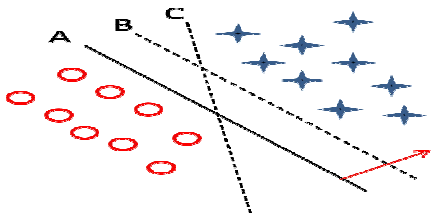

Fig. 3 Using support vector machine on a classification problem

V. PRODUCT ASPECT RANKING

After the aspect sentiment classification, classified aspects are then ranked. Zheng et al [26] described the probabilistic aspect ranking algorithm in their literature.

VI. CONCLUSION

This survey paper presented an overview on the product aspect ranking techniques to identify important aspects of products. Product aspect ranking process contains three main steps i.e. product aspect identification, aspect sentiment classification and aspect ranking. We have conducted a survey which illustrates various methods for aspect identification and sentiment classification.

REFERENCES

1. Ghose and P. G. Ipeirotis,"Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," IEEE Trans. Knowl. Data Eng., vol. 23, no. 10, pp. 1498–1512. Sept. 2010.
2. .L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," J. Mach. Learn., vol. 2, pp. 139–154, Dec. 2011.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

3. M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. SIGKDD, Seattle, WA, USA, 2004, pp. 168–177.
4. Ohana and B. Tierney, "Sentiment classification of reviews using SentiWordNet," in Proc. IT&T Conf., Dublin, Ireland, 2009.
5. 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.
6. Miller G, Beckwith R, Fellbaum C, Gross D, Miller K. WordNet: an on-line lexical database. Oxford Univ. Press; 1990.
7. Hu Minging, Liu Bing. Mining and summarizing customer reviews. In: Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'04); 2004.
8. Kim S, Hovy E. Determining the sentiment of opinions. In: Proceedings of interntional conference on Computational Linguistics (COLING'04); 2004.
9. Mohammad S, Dunne C, Dorr B. Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP'09); 2009.
10. NLProcessor – Text Analysis Toolkit. 2000. http://www.infogistics.com/textanalysis.html
11. Agrawal, R. and Srikant, R. 1994. Fast algorithm for mining association rules. VLDB'94.
12. 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.
13. Joachims T. Probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In: Presented at the ICML conference; 1997.
14. Aizerman M, Braverman E, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. Autom Rem Cont 1964:821–37.
15. Chin Chen Chien, Tseng You-De. Quality evaluation of product reviews using an information quality framework. Decis Support Syst 2011;50:755–68.
16. Li Yung-Ming, Li Tsung-Ying. Deriving market intelligence from microblogs. Decis Support Syst 2013.
17. Miller, G., Beckwith, R, Fellbaum, C., Gross, D., and Miller, K. 1990. Introduction to WordNet: An on-line lexical database. International Journal of Lexicography(special issue), 3(4):235-312.
18. F. Li et al., "Structure-aware review mining and summarization,"in Proc. 23rd Int. Conf. COLING, Beijing, China, 2010, pp. 653–661.
19. A. M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in Proc. HLT/EMNLP, Vancouver, BC,Canada, 2005, pp. 339–346.
20. Q. Su et al., "Hidden sentiment association in chinese web opinion mining," in Proc. 17th Int. Conf. WWW, Beijing, China,2008, pp. 959–968.
21. Y. Wu, Q. Zhang, X. Huang, and L. Wu, "Phrase dependency parsing for opinion mining," in Proc. ACL, Singapore, 2009, pp. 1533–1541.
22. 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.
23. A. Liu, Sentiment Analysis and Opinion Mining. Mogarn & Claypool Publishers, San Rafael, CA, USA, 2012.
24. Kang Hanhoon, Yoo Seong Joon, Han Dongil. Senti-lexicon and improved Naı¨ve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst Appl 2012;39:6000–10.
25. Kaufmann JM. JMaxAlign: A Maximum Entropy Parallel Sentence Alignment Tool. In: Proceedings of COLING'12: Demonstration Papers, Mumbai; 2012. p. 277–88
26. Zheng-Jun Zha, Jianxing Yu, Jinhui Tang, Meng Wang, and Tat-Seng Chua," Product Aspect Ranking and
27. Its Applications", IEEE transactions on knowledge ans data engineering, Vol. 26, No.5, May 2014.
28.

BIOGRAPHY

Rutuja V. Tikait received B.E. degree in Information Technology from Amravati University in 2009. 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.