IJCSEC- International Journal of Computer Science and Engineering Communications.
Vol.3, Issue 2, 2015, Page.695-700, ISSN: 2347–8586

Multiple Aspect Ranking Using Sentiment Classification for Data Mining

N.Anandhi[1] R.Vasanthi[2]

[1]*PG Student* [2]*Asst. Prof Dept. of CSE,*
[1&2]*Affiliated to Anna University Chennai, Dept. of Computer Science and Engineering*
Idhaya Engineering College for Women

nanandhibtech@gmail.com, vasanthiattur@gmail.com

Abstract—Numerous consumer reviews of products are now available on the Internet. Consumer reviews contain rich and valuable knowledge for both firms and users. However, the reviews are often disorganized, leading to difficulties in information navigation and knowledge acquisition. This article proposes a product aspect ranking framework, which automatically identifies the important aspects of products from online consumer reviews, aiming at improving the usability of the numerous reviews. The important product aspects are identified based on two observations: 1) the important aspects are usually commented on by a large number of consumers and 2) consumer opinions on the important aspects greatly influence their overall opinions on the product. We then develop a probabilistic aspect ranking algorithm to infer the importance of aspects by simultaneously considering aspect frequency and the influence of consumer opinions given to each aspect over their overall opinions. The experimental results on a review corpus of 21 popular products in eight domains demonstrate the effectiveness of the proposed approach. Moreover, we apply product aspect ranking to two real-world applications, i.e., document-level sentiment classification and extractive review summarization, and achieve significant performance improvements, which demonstrate the capacity of product aspect ranking in facilitating real-world applications.

Index Terms—Product aspects, aspect ranking, aspect identification, sentiment classification, consumer review, extractive review summarization

1 INTRODUCTION

Recent years have witnessed the rapidly expanding e-commerce. A recent study from ComScore reports that online retail spending reached $37.5 billion in Q2 2011 U.S. Millions of products from various merchants have been offered online. For example, Bing Shopping[1] has indexed more than five million products. Amazon.com archives a total of more than 36 million products. Shopper.com records more than five million products from over 3,000 merchants. Most retail Websites encourage consumers to write reviews to express their opinions on various aspects of the products. Here, an *aspect*, also called *feature* in literatures, refers to a component or an attribute of a certain product. A sample review "*The battery life of Nokia N95 is amazing.*" reveals positive opinion on the aspect "*battery life*" of Product Nokia N95.For example, CNet.com involves more than seven million product reviews; whereas Pricegrabber.com contains millions of reviews on more than 32 million products in 20 distinct categories over 11,000 merchants. Such numerous consumer reviews contain rich and valuable knowledge and have become an important resource for both consumers and firms [9]. Consumers commonly seek quality information from online reviews prior to purchasing a product, while many firms use online reviews as important feedbacks in their product development, marketing, and consumer relationship management. Generally, a product may have hundreds of aspects. For example, *iPhone 3GS* has more than three hundred aspects such as "*usability*," "*design*," "*Application*," "*3G network*." We argue that some aspects are more important than the others, and have greater impact on the eventual consumers' decision making as well as firms' product development strategies.

For example, some aspects of *iPhone 3GS*, e.g., "*usability*" and "*battery*," are concerned by most consumers, and are more important than the others such as "*usb*" and "*button*." For a camera product, the aspects such as "*lenses*" and "*picture quality*" would greatly influence consumer opinions on the camera, and they are more important than the aspects such as "*a/v cable*" and "*wrist strap*." Hence, identifying important product aspects will improve the usability of numerous reviews and is beneficial to both consumers and firms. product aspect ranking framework to automatically identify the important aspects of products from online consumer reviews. Our assumption is that the important aspects of a product possess the following characteristics:(a) they are frequently commented in consumer reviews; and (b) consumers' opinions on these aspects greatly influence their overall opinions on the product. A straightforward frequency-based solution is to regard the aspects that are frequently commented in consumer reviews as important. However, consumers' opinions on the frequent aspects may not influence their overall opinions on the product, and would not influence their purchasing decisions. For example, most consumers frequently criticize the bad "*signal connection*" of *iPhone 4*, but they may still give high overall ratings to *iPhone 4*.

On the contrast, some aspects such as "*design*" and "*speed*," may not be frequently commented, but usually are more important than "*signal connection*." Therefore, the frequency-based solution is not able to identify the truly important aspects. On the other hand, a basic method to exploit the influence of consumers' opinions on specific aspects over their overall ratings on the product is to count the cases where their opinions on specific aspects and their overall ratings are consistent, and then ranks the aspects according to the number of the consistent cases. This method simply assumes that an overall rating was derived from the specific opinions on different aspects individually, and cannot precisely characterize the correlation between the specific opinions and the overall rating. Hence, we go beyond these methods and propose an effective aspect ranking approach to infer the importance of product aspects. As shown in Fig. 1, given the consumer reviews of a particular product, we first identify aspects in the reviews by a shallow dependency parser [37] and then analyze consumer opinions on these aspects via a sentiment classifier. We then develop a probabilistic aspect ranking algorithm, which effectively exploits the aspect frequency as well as the influence of consumers' opinions given to each aspect over their overall opinions on the product in a unified probabilistic model.

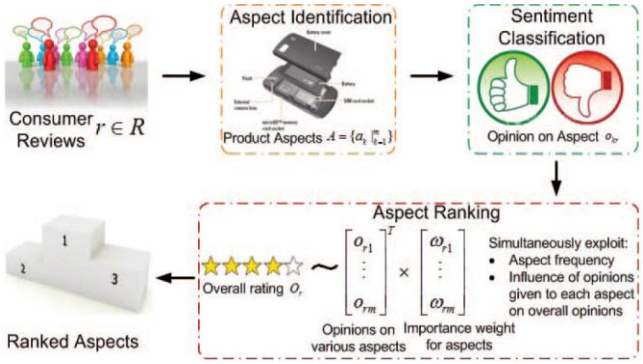

Fig. 1. Flowchart of the product aspect ranking framework.

Product aspect ranking is beneficial to a wide range of real-world applications. In this paper, we investigate its usefulness in two applications, i.e. document-level sentiment classification that aims to determine a review document as expressing a positive or negative overall opinion, and extractive review summarization which aims to summarize consumer reviews by selecting informative review sentences.

Perform extensive experiments to evaluate the efficacy of aspect ranking in these two applications and achieve significant performance improvements. Product aspect ranking was first introduced in our previous work [38]. Compared to the preliminary conference version [38], this article has no less than the following improvements: (a) it elaborates more discussions and analysis on product aspect ranking problem; (b) it performs extensive evaluations on more products in more diverse domains; and (c) it demonstrates the potential of aspect ranking in more real-world applications.

In summary, the main contributions of this article include:

- The propose a product aspect ranking framework to automatically identify the important aspects of products from numerous consumer reviews.
- In develop a probabilistic aspect ranking algorithm to infer the importance of various aspects by simultaneously exploiting aspect frequency and the influence of consumers' opinions given to each aspect over their overall opinions on the product.
- We demonstrate the potential of aspect ranking in real-world applications. Significant performance improvements are obtained on the applications of document-level sentiment classification and extractive review summarization by making use of aspect ranking.

2. PRODUCT ASPECT RANKING FRAMEWORK

The details of the proposed Product Aspect Ranking framework. We start with an overview of its pipeline (see Fig. 1) consisting of three main components: (a) aspect identification; (b) sentiment classification on aspects; and (c) probabilistic aspect ranking. Given the consumer reviews of a product, we first identify the aspects in the reviews and then analyze consumer opinions on the aspects via a sentiment classifier. Finally, we propose a probabilistic aspect ranking algorithm to infer the importance of the aspects by simultaneously taking into account aspect frequency and the influence of consumers' opinions given to each aspect over their overall opinions.

2.1 Product Aspect Identification:

Consumer reviews are composed in different formats on various forum Websites. The Websites such as *CNet.com* require consumers to give an overall rating on the product, describe concise positive and negative opinions (i.e. *Pros and Cons*) on some product aspects, as well as write a paragraph of detailed review in free text. Some Websites, e.g., *Viewpoints.com*, only ask for an overall rating and a paragraph of free-text review. The others such as *Reevoo.com* just require an overall rating and some concise positive and negative opinions on certain aspects. In summary, besides an overall rating, a consumer review consists of *Pros* and *Cons* reviews, free text review, or both. For the *Pros* and *Cons* reviews, we identify the aspects by extracting the frequent noun terms in the reviews. For identifying aspects in the free text reviews and In order to obtain more precise identification of aspects, we here propose to exploit the *Pros* and *Cons* reviews as auxiliary knowledge to assist identify aspects in the free text reviews. In particular, we first split the free text reviews into sentences, and parse each sentence using Stanford parser2. The frequent noun phrases are then extracted from the sentence parsing trees as candidate aspects. Since these candidates may contain noises, we further leverage the *Pros* and *Cons* reviews to assist identify aspects from the candidates. We collect all the frequent noun terms extracted from the *Pros* and *Cons* reviews to form a vocabulary. We then represent each aspect in the *Pros* and *Cons* reviews into a unigram feature, and utilize all the aspects to learn a one-class Support Vector Machine (SVM) classifier. The resultant classifier is in turn used to identify aspects in the candidates extracted from the free text reviews. As the identified aspects may contain some synonym terms, such as "*earphone*" and "*headphone*," we perform synonym clustering to obtain unique aspects. In particular, we collect the synonym terms of the aspects as features. The synonym terms are collected from the synonym dictionary Website3.

2.2. Sentiment Classification on Product Aspects

The task of analyzing the sentiments expressed on aspects is called aspect-level sentiment classification in literature. Exiting techniques include the supervised learning approaches and the lexicon-based approaches, which are typically unsupervised. The lexicon-based methods utilize a sentiment lexicon consisting of a list of sentiment words, phrases and idioms, to determine the sentiment orientation on each aspect. While these method are easily to implement, their performance relies heavily on the quality of the sentiment lexicon.

On the other hand, the supervised learning methods train a sentiment classifier based on training corpus. The classifier is then used to predict the sentiment on each aspect. Many learning-based classification models are applicable, for example, Support Vector Machine (SVM), Naive Bayes, and Maximum Entropy (ME) model etc. Supervised learning is dependent on the training data and cannot perform well without sufficient training samples. However, labeling training data is labor intensive and time-consuming. In this work, the *Pros* and reviews (i.e., positive samples) and *Cons* reviews (i.e., negative samples). The classifier can be SVM, Naïve Bayes or Maximum Entropy model. Given a free text review that may cover multiple aspects, we first locate the opinionated expression that modifies the corresponding aspect, e.g. locating the expression "*well*" in the review "*The battery of Nokia N95 works well.*" for the aspect "*battery.*" A sentiment classifier is then learned from the *Pros* reviews (i.e., positive samples) and *Cons* reviews (i.e., negative samples). The classifier can be SVM, Naïve Bayes or Maximum Entropy model . Given a free text review that may cover multiple aspects, we first locate the opinionated expression that modifies the corresponding aspect, e.g. locating the expression "*well*" in the review "*The battery of Nokia N95 works well.*" for the aspect "*battery.*" Generally, an opinionated expression is associated with the aspect if it contains at least one sentiment term in the sentiment lexicon, and it is the closest one to the aspect in the parsing tree within the context distance of 5. The learned sentiment classifier is then leveraged to determine the opinion of the opinionated expression, i.e. the opinion on the aspect.

2.3. Probabilistic Aspect Ranking Algorithm

In this section, a probabilistic aspect ranking algorithm to identify the important aspects of a product from consumer reviews. Generally, important aspects have the following characteristics: (a) they are frequently commented in consumer reviews; and (b) consumers' opinions on these aspects greatly influence their overall opinions on the product. The overall opinion in a review is an aggregation of the opinions given to specific aspects in the review, and various aspects have different contributions in the aggregation. That is, the opinions on (un)important aspects have strong (weak) impacts on the generation of overall opinion. To model such aggregation, we formulate that the overall rating *Or* in each review *r* is generated based on the weighted sum of the opinions on specific aspects.

3. EVALUATIONS

We conduct extensive experiments to evaluate the effectiveness of the proposed product aspect ranking framework, including product aspect identification, sentiment classification on aspects, and aspect ranking.

3.1 Experimental Data and Settings

The details of our product review corpus, which is publicly available by request. This dataset contains consumer reviews on 21 popular products in eight domains. There are 94,560 reviews in total and around 4,503 reviews for each product on average. These reviews were crawled from multiple prevalent forum Websites, including *cnet.com*, *viewpoints.com*, *reevoo.com*, *gsmarena.com* and *pricegrabber.com*. The reviews were posted between June 2009 and July 2011. Eight annotators were invited to annotate the ground truth on these reviews.

3.2 Evaluations of Product Aspect Identification on Free Text Reviews

Our aspect identification approach with the following two methods: (a) the method proposed by Hu and Liu in [12], which extracts nouns and noun phrases as aspect candidates, and identifies aspects by rules learned from association rule mining; and (b) the method proposed by Wu *et al*. in [37], that extracts noun phrases from a dependency parsing tree as aspect candidates, and identifies aspects by a language model built on the reviews.

3.3 Evaluations of Sentiment Classification on Product Aspects

The following methods of sentiment classification: (a) one unsupervised method.The opinion on each aspect is determined by referring to the sentiment lexicon *SentiWordNet* [23]. This lexicon contains a list of positive/negative sentiment words. The opinionated expression modifying an aspect is classified as positive (or negative) if it contains a majority of words in the positive (or negative) list; and (b) three supervised methods. We employed three supervised methods proposed in Pang *et al*. [25], including Naïve Bayes (**NB**), Maximum Entropy (**ME**), and Support Vector Machine (**SVM**). The sentiment classifiers were trained on the *Pros* and *Cons* reviews.

3.4 Evaluations of Aspect Ranking

In order to evaluate the effectiveness on aspect ranking, we compared the proposed aspect ranking algorithm with the following three methods: (a) **Frequency-based method**, which ranks the aspects according to aspect frequency; (b) **Correlation-based method**, which measures the correlation between the opinions on specific aspects and the overall ratings. It ranks the aspects based on the number of cases when such two kinds of opinions are consistent; and (c) **Hybrid method**, that captures both aspect frequency and the correlation by a linear combination, as $\lambda \cdot$ *Frequency-based Ranking* + $(1-\lambda) \cdot$ *Correlation-based Ranking*, where λ is set to 0.5 in the experiments.

4 .APPLICATIONS

Aspect ranking is beneficial to a wide range of real world applications. We here investigate its capacity in two applications, i.e. document-level sentiment classification on review documents, and extractive review summarization.

4.1 Document-level Sentiment Classification

The goal of document-level sentiment classification is to determine the overall opinion of a given review document. A review document often expresses various opinions on multiple aspects of a certain product. The opinions on different aspects might be in contrast to each other, and have different degree of impacts on the overall opinion of the review document. For example, a sample review document of *iPhone 4* .It expresses positive opinions on some aspects such as "*reliability*," "*easy to use*," and simultaneously criticizes some other aspects such as "*touch screen*," "*quirk*," "*music play*." Finally, it assigns an high overall rating (i.e., positive opinion) on *iPhone 4* due to that the important aspects are with positive opinions. Hence, identifying important aspects can naturally facilitate the estimation of the overall opinions on review documents. This observation motivates us to utilize the aspect ranking results to assist document-level sentiment classification.

4.2 Extractive Review Summarization

As aforementioned, for a particular product, there is an abundance of consumer reviews available on the internet. However, the reviews are disorganized. It is impractical for user to grasp the overview of consumer reviews and opinions on various aspects of a product from such enormous reviews. On the other hand, the Internet provides more information than is needed. Hence, there is a compelling need for automatic review summarization, which aims to condense the source reviews into a shorter version preserving its information content and overall meaning. Existing review summarization methods can be classified into abstractive and extractive summarization. An abstractive summarization attempts to develop an understanding of the main topics in the source reviews and then express those topics in clear natural language. It uses linguistic techniques to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. An extractive method summarization method consists of selecting important sentences and paragraphs etc. from the original reviews and concatenating them into shorter from.

5. RELATED WORKS

The two evaluated real-world applications. We start with the works on aspect identification. Existing techniques for aspect identification include supervised and unsupervised methods. Supervised method learns an extraction model from a collection of labeled reviews. The extraction model, or called extractor, is used to identify aspects in new reviews. Most existing supervised methods are based on the sequential learning (or sequential labeling) technique [18]. For example, Wong and Lam [36] learned aspect extractors using *Hidden Markov Models* and *Conditional Random Fields*, respectively. Jin and Ho [11] learned a lexicalized HMM model to extract aspects and opinion expressions, while Li *et al*. [16] integrated two CRF variations, i.e., Skip-CRF and Tree-CRF. All these methods require sufficient labeled samples for training. However, it is time-consuming and labor-intensive to label samples. On the other hand, unsupervised methods have emerged recently. The most notable unsupervised approach was proposed by Hu and Liu [12]. They assumed that product aspects are nouns and noun phrases. The approach first extracts nouns and noun phrases as candidate aspects. The occurrence frequencies of the nouns and noun phrases are counted, and only the frequent ones are kept as aspects. Subsequently, Popescu and Etzioni [28] developed the *OPINE* system, which extracts aspects based on the *KnowItAll* Web information extraction system [8]. Mei *et al*. [22] utilized a probabilistic topic model to capture the mixture of aspects and sentiments simultaneously. Su *et al*. [32] designed a mutual reinforcement strategy to simultaneously cluster product aspects and opinion words by iteratively fusing both content and sentiment link information. Recently, Wu *et al*. [37] utilized a phrase dependency parser to extract noun phrases from reviews as aspect candidates.

6 .CONCLUSION

The product aspect ranking framework to identify the important aspects of products from numerous consumer reviews. The framework contains three main components, i.e., product aspect identification, aspect sentiment classification, and aspect ranking. The algorithm simultaneously explores aspect frequency and the influence of consumer opinions given to each aspect over the overall opinions. The product aspects are finally ranked according to their importance scores. Moreover applied product aspect ranking to facilitate two real-world applications, i.e., document-level sentiment classification and extractive review summarization.

REFERENCES

[1] Bezdek.J.C et al, (2003) “Convergence of alternating optimization,” J. Neural Parallel Scientific Comput., vol. 11, no. 4, pp. 351–368.

[2] Ding.X et al, (2008) “A holistic lexicon-based approach to opinion mining,” in Proc. WSDM, New York, NY, USA, pp. 231–240.

[3] Gupta.V et al, (2010) “A survey of text summarization extractive techniques,” J. Emerg. Technol. Web Intell., vol. 2, no. 3, pp. 258–268.

[4] Ghose.A et al, (2010) “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.

[5] Paltoglou.G et al, (2010) “A study of information retrieval weighting schemes for sentiment analysis,” in Proc. 48th Annu. Meeting ACL, Uppsala, Sweden, pp. 1386–1395.

[6] Pang.B et al, (2002) “Thumbs up? Sentiment classification using machine learning techniques,” in Proc. EMNLP, Philadelphia, PA, USA, pp. 79–86.8.

[7] Pang.B et al, (2004) “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts techniques,” in Proc. ACL, Barcelona, Spain, pp. 271–278.

[8] Pang.B et al, (2008) “Opinion mining and sentiment analysis,” in Found. Trends Inform. Retrieval, vol. 2, no. 1–2, pp. 1–135.

[9] Snyder.B et al, (2007) “Multiple aspect ranking using the good grief algorithm,” in Proc. HLT-NAACL, New York, NY, USA, pp. 300–307.

[10] Zheng-Jun Zha et al, (2014) “Product aspect Ranking and its Application”.