PRODUCT ASPECT RANKING AND ITS APPLICATIONS USING SPLSA

Anuradha.P ., Sujatha, T and M.Vetripriya

Dept. of Information Technology, Madha Engineering College, Kundrathur, Chennai-69.

ABSTRACT

Nowadays user reviews has become an increasingly popular way for people to express opinions and sentiments toward the products bought or services received. Analyzing the large volume of online reviews available would produce useful actionable knowledge that could be of economic values to vendors and other interested parties. The reviews are extracted from various websites and the fake or spam reviews are identified and then eliminated. The product aspects are identified by Phrase Dependency parser. The sentiment Probabilistic Latent Semantic Analysis has been used for the analysis of sentiment of the reviews. Probabilistic aspect ranking algorithm is used to rank the aspects by inferring the importance of aspects by simultaneously considering aspect frequency and the influence of consumer opinions given to each aspect over their overall opinions.

Keywords: Opinion mining, S-PLSA, ARSA, Review mining , sentiment analysis , aspect identification, opinion spamming.

I.INTRODUCTION

With the rapid expansion of e-commerce, more and more products are sold on the Web, and people are also buying products online. In order to enhance customer satisfaction and shopping experience, it has become a common practice for online merchants to enable their customers to review or to express opinions on the products. With more users becoming comfortable with the Web, an increasing number of people are writing reviews. As a result, the number of reviews that a product receives is high. Some popular products can get hundreds of reviews at some large merchant sites [1]. Furthermore, many reviews are long and have only a few sentences

containing opinions on the product. This makes it hard for a potential customer to read them to make an informed decision on whether to purchase the product or not. If he/she only reads a few reviews, he/she may get a biased view. The large number of reviews also makes it hard for product manufacturers to keep track of customer opinions of their products. The Sentiment Analysis is an application of Natural Language Processing which targets on the identification of the sentiment (positive vs negative vs neutral), the subjectivity (objective vs subjective) and the emotional states of the document. Sentiment classification is an opinion mining activity concerned with determining the overall sentiment orientation of the opinions contained within a given document. It is assumed in general that the document being inspected contains subjective information, as in product reviews .The Sentiment Probabilistic Latent Semantic Analysis (SPLSA) has been used for the analysis of the sentiments in the reviews, which focuses on all the statement in the reviews. Opinionated social media such as product reviews are now widely used by individuals and organizations for their decision making. However, due to the reason of profit or fame, people try to game the system by opinion spamming (e.g., writing fake reviews) to promote or demote some target products [2].

II. RELATED WORK

The wide spread use of online reviews as a way of conveying views and comments has provided a unique opportunity to understand the general public's sentiments and derive business intelligence. With the rapid growth of online reviews, review mining has attracted a great deal of attention. Early work in this area was primarily focused on determining the semantic orientation of reviews. Sentiment Probabilistic Latent Semantic Analysis is a novel statistical technique for the analysis of two mode and co-occurrence data, which has applications in information retrieval and filtering, natural language processing, machine learning from text, and in related areas. Compared to standard Latent Semantic Analysis which starts from linear algebra and performs a Singular Value Decomposition of co-occurrence tables, the proposed method is based on a mixture decomposition derived from a latent class model. This results in a more principled approach which has a solid foundation in statistics.

Marketing plays an important role in the newly released products, customer word of mouth can be a crucial factor that determines the success in the long run, and such effect is largely magnified thanks to the rapid growth of Internet. Therefore, online product reviews can be very

valuable to the vendors in that they can be used to monitor consumer opinions toward their products in real time, and adjust their manufacturing, servicing, and marketing strategies accordingly.Some studies attempt to answer the question of whether the polarity and the volume of reviews available online have a measurable and significant effect on actual customer purchasing [3].

Compared to sentiment mining, identifying the quality of online reviews has received relatively less attention. A few recent studies along this direction attempt to detect the spam that exists in online reviews. Jindal and Liu present a categorization of review spams, and propose some novel strategies to detect different types of spams [4]. Liu et al. propose a classification-based approach to discriminate the fake reviews from others, in the hope that such a filtering strategy can be incorporated to enhance the task of opinion summarization [5]. Elkan develops a complete framework that consists of six different components, for retrieving and filtering online documents with uneven quality.

Opinionated social media such as product reviews are now widely used by individuals and organizations for their decision making. However, due to the reason of profit or fame, people try to game the system by opinion spamming (e.g., writing fake reviews) to promote or demote some target products. For reviews to reflect genuine user experiences and opinions, such spam reviews should be detected. Opinion spam focused on detecting fake reviews and individual fake reviewers [6].

Recently, information extraction from texts was studied by several researchers. Their focus is on using machine learning and NLP methods to extract/classify named entities and relations [7]. In order to analyze reviews, the reviews are to be extracted from Web pages. Note that this step is not needed if a merchant who already has reviews at its site (e.g., amazon.com) or a dedicated review site (e.g., epinions.com) wants to provide the service.

Example:

My SLR is on the shelf

By shortstop24, Aug 09 ‘03

Pros: Great photos, easy to use, good manual, many options, takes videos

Cons: Battery usage, included software could be improved, included 16MB is stingy.

I had never used a digital camera prior to purchasing the Canon A70. I have always used a SLR.

Pros in the above example can be separated into 5 segments.

great photos <photo>

easy to use <use>

good manual <manual>

many options <option>

takes videos <video>

Cons in the above example can be separated into 3 segments:

battery usage <battery>

included software could be improved <software>

included 16MB is stingy <16MB>⇒<memory>

The pros reflect the positive opinion about the reviews of the product. The cons reflect the negative opinions or the review that helps for the improvement in the product which helps to increase the production of the product for further sales. To perform the extraction task automatically is a non-trivial task. Manually browsing the Web and doing cut-and-paste is clearly not acceptable. It is also too time consuming to write a site specific extraction program for each site. Fortunately, there are existing technologies for this purpose. One approach is wrapper induction [8]. A wrapper induction system allows the user to manually label a set of reviews from each site and the system learns extraction rules from them. These rules are then used to automatically extract reviews from other pages at the same site. Another approach is to automatically find patterns from a page that contains several reviews. These patterns are then employed to extract reviews from other pages of the site. Both these approaches are based on the fact that reviews at each site are displayed according to some fixed layout templates.

III. EXISTING SYSTEM

It extract user reviews from websites such as amazon.com and cnet.com, then the aspects are identified using stanford parser and the sentiment analysis is done with the machine learning techniques such as PLSA. Then, ranking is done with Probabilistic aspect ranking algorithm. The use of PLSA has some drawbacks such as, it has a single dataset and so it has to traverse the

whole dataset in order to find a single aspect. It focuses on all the comments present in the reviews which are difficult to analyze the sentimental words. Apart from this drawback, there are other drawbacks like, the reviews can also be spam or fake, and they should be identified and eliminated for the ranking to be accurate.

IV. PROPOSED SYSTEM

Gathering and comparing consumer opinions of competing products from the Web for marketing intelligence and for product benchmarking is an important problem. The fake and spam reviews are detected and eliminated from the extracted reviews. The consumer reviews are collected from websites like amazon.com, cnet.com etc. Phrase dependency parser is used for aspect identification. The analyzation of the sentiment is done using SPLSA. The important aspects are ranked based on Probabilistic Aspect Ranking algorithm.

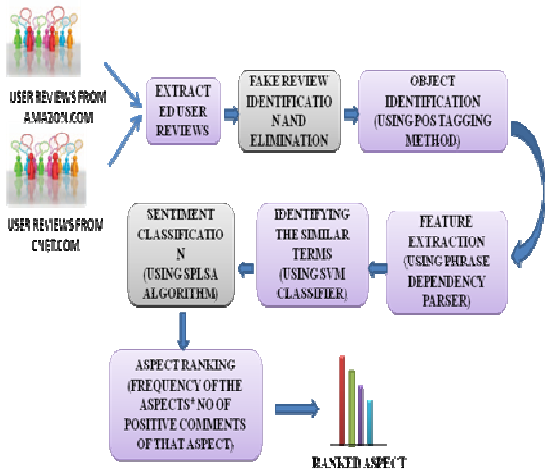

Figure 1: Architecture of proposed system

The techniques in proposed system are:

1. Opinion spamming
2. Phrase dependency parser
3. S-PLSA
4. Probabilistic Aspect Ranking Algorithm

Opinion spamming

Fake reviews are untruthful reviews that are written with hidden motives. They often contain undeserving positive opinions about some target entities (products or services) in order to promote the entities and to create negative opinions about some other entities in order to damage

their reputations. These reviews do not comment on the specific products or services that they are supposed to review, but only comment on the brands or the manufacturers of the products. There are two features with which the fake reviews can be easily identified [10].

(i)Maximum Number of Reviews: Posting many reviews in a single day reflects abnormal reviewing pattern and can be used as a behavioral feature. This feature simply computes the maximum number of reviews posted in a day for an author. It is normalized by the maximum value in the dataset.

(ii)Extreme Rating: Opinion spamming typically projects entities incorrectly either in a very positive or a very negative light [11]. Thus, on a 5-star rating scale, spammers are likely to give extreme ratings (1 or 5 stars) in order to promote or to demote entities. The following review feature accounts whether the associated star rating of the review was extreme or not.

Example:

Posted by: Sam on December 2, 2011

Ratings: "The canon G12 camera is superb. The canon G12 camera has good lifetime. My sister loved canon G12 camera. The canon G12 camera has perfect lens".

Phrase dependency parser: A phrase dependency parser to extract noun phrases, which form candidate aspects [12]. To filter out the noises, we use a language model by an intuition that the more likely a candidate to be an aspect, the more closely it related to the reviews. The language model was built on product reviews, and used to predict the related scores of the candidate aspects. The candidates with low scores were then filtered out. However, such language model might be biased to the frequent terms in the reviews and cannot precisely sense the related scores of the aspect terms, as a result cannot filter out the noises effectively. In order to obtain more precise identification of aspects, we here propose to exploit the Pros and Cons reviews as auxiliary knowledge to assist identifies aspects in the free text reviews. We then use SVM classifier [12] in order to cluster the similar terms (i.e.) candidate aspects.

Example: We really enjoyed using the SD500 Canon Power Shot.

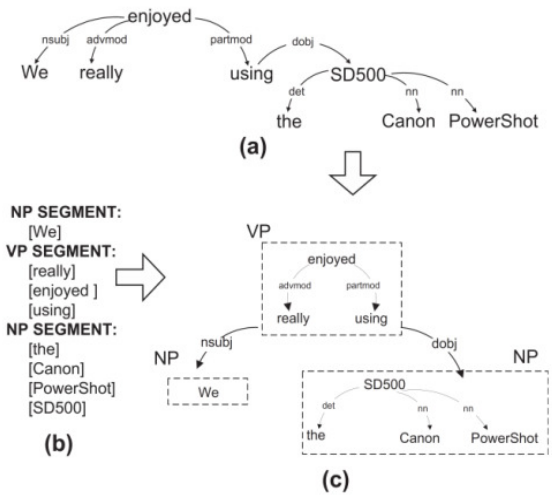

Figure 2: Example of Phrase Dependency Parsing

S-PLSA - **S-PLSA** (Sentimental PLSA) technique is used for the purpose of sentiment classification [1]. In this technique, instead of taking a "bag-of-words" approach and considering all the words present in the blogs, it is enough to focus on the words that are sentiment related. The Sentiment has been classified into "good", "spam" and "bad". The corresponding sentimental words are stored in the respective database. Some of other related works are Review mining, that is nothing but of collecting all the reviews that are commented in the review space by the reviewer and the customer about the product.

In a similar way, an explicit feature and an implicit feature can be easily identified in a sentence. For example, "battery life" in the following two opinion sentences/segments is an explicit feature: "The battery life of this camera is too short", "Battery life too short", "Size" is an implicit feature in the following two opinion sentences as it does not appear in each sentence but it is implied: "This camera is too large", "Too big"

As per the review, the rating has been given which is very effective for the product manufacturer to improve their product quality. Instead of considering all the words present in the blogs, we focus primarily on the words that are sentiment related which can reduce the processing time. This is the main advantage of SPLSA over PLSA.

Example: "The battery life is good".

Here the **PLSA [8]** takes the whole statement as a sentiment. But the **S-PLSA [1]** takes the word "**good**" as a sentimental word.

Probabilistic aspect ranking algorithm: A probabilistic aspect ranking algorithm is used 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.

Aspect Ranking = (Frequency of Aspects) *(Number Of Positive Comments Of that Aspect).

Example : Battery life (Frequency count=7), (Positive comments=3)
Picture quality (Frequency count=9), (Positive comments=6)
Audio clarity (Frequency count=6), (Positive comments=4)

RANKING:	**RANKED ASPECTS:**
Battery life= (7*3) = 21	1.Picture quality
Picture quality= (9*6) = 54	2.Audio clarity
Audio clarity= (6*4) = 24	3. Battery life

V. CONCLUSION

Product aspect ranking with review quality framework is used to identify the important aspects of products from numerous consumer reviews. The framework contains four main components, i.e., fake review identification and elimination, product aspect identification, aspect sentiment classification, and aspect ranking. The Pros and Cons reviews are first identified to improve aspect identification and sentiment classification on free-text reviews. Since the reviews may also be fake, we identified those fake reviews and eliminated it to reduce the processing time. Aspects are identified by phrase dependency parser which uses POS tagging to identify the noun

and verb phrases and sentiment classification is done by SPLSA, which is the means of "summarizing" sentiment information from reviews and also the accuracy and effectiveness of aspect ranking is improved using aspect ranking algorithm. For future enhancement, to further improve the accuracy of prediction by considering the quality factor, with a focus on predicting the quality of a review in the absence of user-supplied indicators, and an ARSQA [Auto regressive Sentiment and Quality Aware model] can be used, an, to utilize sentiments and quality for predicting product sales performance.

VI. REFERENCES

1. Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang and Aijun An, Membesr, IEEE," Mining Online Reviews for Predicting Sales Performance: A Case Studyin the Movie Domain", National Natural Science Foundation of China Grants, 24 (4): 1-5.
2. B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," Proc. 14th Int'l Conf. World Wide Web (WWW), pp. 342-351, 2005.
3. D.Gruhl, R.Guha, R.kumar, J.novak, and A.Tomkins, "The Predictive Power of Online chatter," Proc.11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining(KDD), pp. 78-87,2005.
4. A.Ghose and P.G. Ipeirotis, "Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews," Procs Ninth Int'l Conf. Electronic Commerce (ICEC), pp. 303-301, 2007.
5. Y.Liu, X. Huang, A. An, and X. Yu, "ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 607-614, 2007.
6. Y.Liu, X. Yu, X. Huang, and A. An, "Blog Data Mining: The Predictive Power of Sentiments," Data Mining for Business Application, pp. 183-195, Springer, 2009.
7. D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information Diffusion through Blogspace," Proc. 13th Int'l Conf. World Wide Web (WWW), pp. 491-501, 2004.
8. N. Archak, A. Ghose, and P.G. Ipeirotis, "Show Me the Money!: Deriving the Pricing Power of Product Features by Mining Consumer Reviews," Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 56 65, 2007.

9. J.A. Chevalier and D. Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," J. Marketing Research, vol. 43, no. 3, pp. 345-354, Aug. 2006.
10. S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," J. Political Economy, vol. 82, no. 1, pp. 34-55, 1974.
11. B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-Based Collaborative Filtering Recommendation Algorithms," Proc. 10th Int'l Conf. World Wide Web (WWW), pp. 285-295, 2001.
12. L. M. Manevitz and M. Yousef. 2002. One-class SVMs for Document Classification. *In Journal of Machine Learning, 2: 139-154.*