Mining Features and Ranking Products From Online Customer Reviews

T.Saranya
Assistant Professor
Department of Computer Science and Engineering
Ranipettai Engineering College

***Abstract*— Now days, E-commerce systems have become extremely important. Large numbers of customers are choosing online shopping because of its convenience, reliability, and cost. As a result, an increasing number of customers post product reviews at merchant websites and express their opinions and experiences in any network space such as Internet forums, discussion groups, and blogs. So there are a large amount of data records related to products on the Web, which are useful for both manufacturers and customers. But it is difficult for customers to make purchasing decisions based on only pictures and short product descriptions. On the other hand, mining product reviews has become a hot research topic and prior researches are mostly based on pre-specified product features to analyze the opinions. In the Existing system, ranking mechanisms typically rank products based on their overall quality and subset of feature phrases mentioned manually. The process of identifying features manually is both time-consuming and may also lead to some features being missed out. In this system, it is proposed to identify the features automatically using online customer reviews. This is expected to yield accurate rankings for each product when compared to the existing system. Thus is expected to perform better.**

Index Terms— Feature Extraction, Sentences Sentiment Orientation, Prank Algorithm.

1. INTRODUCTION

As Internet has become a part of people's daily lives, e-commerce has also increased day by day. More products are sold on the Web and people are purchasing through the Web. In order to share their shopping experiences and feedbacks, an increasing number of customers post product reviews at merchant websites and express their opinions and experiences in any network space such as internet forums, discussion groups and blogs, which has a great wealth of opinion about products. It is becoming increasingly difficult for customers to make purchasing decisions based on only pictures and short product descriptions. So in order to avoid confusion, ranking of each product can be made from the features automatically given by customer reviews. A systematic procedure of opinion mining is formed in many researchers' effort, and Hu and Liu in [8] propose one useful method which is feature-based opinion summarization of reviews.

Opinion mining involves structured summarization not about a free document. It involves mining the features of the product that customers have expressed opinions on and then ranking the features according to their frequencies that they appear in the reviews. Users are usually focused on the product features that customers have positive or negative opinions on. Extracting product features are the fundamental step of opinion mining. For each feature, opinion sentences will be identified in each review and each opinion sentence's semantic orientation will be determined. The specific reviews for the products by the potential customers and summarized results are analyzed generated using the discovered information. Previous ranking mechanisms typically, rank products based on their overall quality and Product Feature based Ranking.

A product has usually multiple product features, each of which plays a different role. Different customers may be interested in different features of a product. Traditionally, many customers have used expert rankings which rate limited a number of products. By using the information obtained from customer reviews, the relationships among products can be modeled constructing a weighted and directed graph.

This paper proposes to implement feature based ranking technique on Customer Online Reviews and to enhance the existing system having the following considerations: (1) To Identify the Features automatically, so as to get more features about each product and cover all features of that product. (2)For the customer, choosing a better product becomes much easier and also more time is saved.

The remainder of this paper is organized as follows: Section 2 describes about the detailed literature review where various techniques available for mining product features from online reviews are outlined and some of the drawbacks of the existing approaches are quoted. Section 3 consists of the details of the proposed system and it depicts the system's structure Section 4 consists of the overall modules involved in the implementation and the performance evaluation of the system. Section 5 presents the results and discussion.

2. LITERATURE REVIEW

In [8], the authors have discussed about mining opinion features in customer reviews. There are many techniques for opinion features, and they can be broadly classified into POS tagging and frequent & infrequent features extraction. It is always possible to detect features by feature extraction, as frequent features only differ from infrequent features because of their opinions. Many of these customer opinions are used to create number of features about product. Feature pruning aims to remove these incorrect features. We present two types of pruning, compact pruning and redundancy pruning are discussed in [8].

In [2], the authors suggest that the Web contains a wealth of opinions about products, politicians, and more, which are expressed in newsgroup posts, review sites, and elsewhere. As a result, the problem of "opinion mining" has seen increasing attention over the last three years from and many others. This paper focuses on product reviews, Product reviews on Web sites such as amazon.com often associate meta-data with each review indicating how positive (or negative) it is using a 5-star scale, and also rank products by how they fare in the reviews at the site. The problem of review mining can be decomposed into the following subtasks 1) Identify product features 2) Identify opinions regarding product features 3) Determine the polarity of opinions.4) Rank opinions based on their strength.

In [14], the authors observe that over the past few years, many researchers studied the problem which is called opinion mining or sentiment analysis to find product features that have been commented on by reviewers and decide whether the comments are positive or negative. The authors focus on the task of identifying the semantic orientations of opinion expressed on each product feature by each reviewer. Semantic orientation means whether the opinion is positive, negative or neutral. It deals with many special words, phrase and language constructs which are based on their linguistic patterns. The method handles implicit features represented by feature indicators and also explicit features.

In [11], the authors suggest identifying comparative sentences in text documents. Comparisons are one of the most convincing ways of evaluation. Extracting comparative sentences from text is useful for many applications. Comparative sentences are classified into different categories based on existing linguistic research. They are also expanded with additional categories that are important in practice. The authors propose a novel approach based on pattern discovery and supervised learning to identify comparative sentences. The basic idea of the technique is to first use a keyword strategy to achieve a high recall and then build a machine learning model to automatically classify each sentence into one of the two classes, "comparative" and "non comparative", based on the filtered data to improve the precision. In building the learning model, class sequential rules automatically generated from the data are used as features.

In [1], a ranking is performed of products based on customer reviews they have received. Consumer reviews of a product are considered more honest, unbiased and comprehensive than a description provided by the seller. This can be done using

1) Subjective sentence 2) Comparative sentence. Also the number of consumer reviews available has increased to an extent where it is no longer possible for a user to peruse them all manually. In this work, the aim is to perform a ranking of products based on customer reviews they have received. Consumer reviews of a product are considered more honest, unbiased and comprehensive than a description provided by the seller.

In [6], a feature-based product ranking technique that mines thousands of customer reviews has been used. First product features are identified within a product category and their frequencies and relative usage are analyzed. For each feature, subjective and comparative sentences in reviews are identified. Sentiment orientations are assigned to these sentences. By using the information obtained from customer reviews, the relationships among products is modeled by constructing a weighted and directed graph. This graph is mined to determine the relative quality of products.

A detailed study of the literature reveals the techniques that can be used for feature extraction and product ranking

3. METHODOLOGY

This section, gives an architectural overview for our product feature extraction system which finally gives the ranking Figure 1.The system performs the extraction in three main steps: product feature extraction, sentence labeling using categories such as subjective, comparative and finally a ranking methodology is used. The inputs to the system are all the reviews of the product in the database. Automatic detection of the product features from a customer review using a feature extraction process is being used in the proposed system in order to overcome the problem of the existing system where the features are manually specified.

Initially all the reviews are collected and loaded into the in the review database. The feature extraction and ranking function are the focus in this paper, a keyword matching strategy is used to identify and tag product features in sentences and extract the product features including explicit feature and implicit feature. The results are filtered via pruning irrelevant features. Next sentence labeling is done. Finally different strategies are used to assign sentiment orientation to sentences and ranking method using Prank algorithm. Below, each of the functions in features Extraction, Sentence labeling and ranking methodology is discussed below in detail.

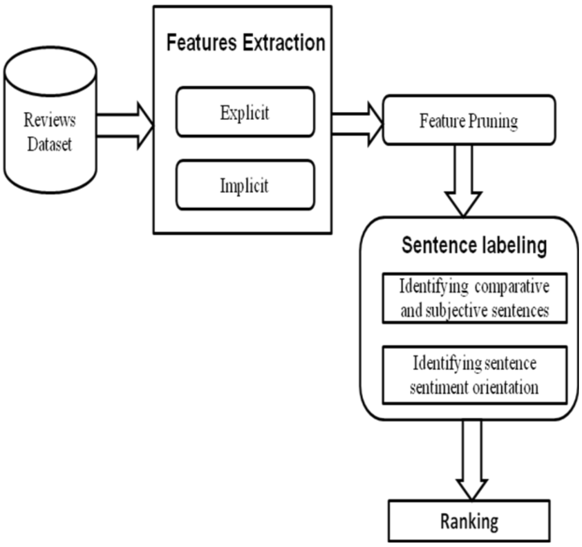

Reviews Dataset
Features Extraction
Explicit
Implicit
Feature Pruning
Sentence labeling
Identifying comparative and subjective sentences
Identifying sentence sentiment orientation
Ranking

Figure.1. System Diagram for Proposed Product Ranking system.

4. IMPLEMENTATION

4.1Review collection and pre-processing

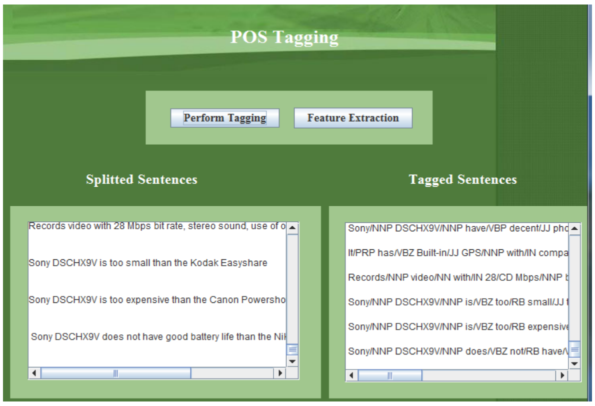
In this module, customer reviews are collected from online retailers like Amazon.com and stored in the database. The collected reviews are tagged with part of speech information.

Part-of-Speech (POS) Tagging

POS tagging is the part-of-speech tagging [13] from natural language processing. The Stanford POS tagger is used which parses each sentence and yields the part-of-speech tag of each word (whether the word is a noun, verb, adjective,etc) and identifies simple noun and verb groups (syntactic chunking). The following shows a sentence with the POS tags.

It/PRP has/VBZ decent/JJ photo/NN quality/NN for/IN midsize/JJ print/NN

For instance, photo/NN indicates a noun. Each sentence is saved in the review database along with the POS tag information of each word in the sentence. The output for the implementation of collecting customer reviews and POS tagging is given in Figure.2.



Records video with 28 Mbps bit rate, stereo sound, use of o

Sony DSCHX9V is too small than the Kodak Easyshare

Sony DSCHX9V is too expensive than the Canon Powersho

Sony DSCHX9V does not have good battery life than the Nik

Sony/NNP DSCHX9V/NNP have/VBP decent/JJ pho

It/PRP has/VBZ Built-in/JJ GPS/NNP with/IN compa

Records/NNP video/NN with/IN 28/CD Mbps/NNP b

Sony/NNP DSCHX9V/NNP is/VBZ too/RB small/JJ t

Sony/NNP DSCHX9V/NNP is/VBZ too/RB expensive

Sony/NNP DSCHX9V/NNP does/VBZ not/RB have/V

Figure.2. Implementation output of collecting customer reviews and POS tagging.

4.2 Product Features Extraction

The feature Extraction phase of the system is developed in order to find features that people are most interested in. The features are classified into explicit and implicit.

4.2.1) Explicit features.

Explicit features are the product features that appear in Noun form in the reviews. Most of the features are expressed explicitly in the reviews. So the focus is on extracting the majority of the explicit features which means finding most of the features in the reviews. Based on the previous step, part of speech tag information and the nearby principle is used to extract the nouns that are described by the adjective. Each noun in the same sentence should appear nearby, and then the noun phrases can be considered as the candidates of product features. One adjective only can be used for one object.

This product has better lens, but the flash is bad
Panasonic has good battery life
It has good quality of the image

Here 'lens' and 'flash' will be treated as two features, because they are not nearby. In sentence 2, 'battery' and 'life' are appearing nearby, so they are considered to form one feature. In sentence 3, 'quality' and 'image' separated by the preposition are also considered nearby and describe one feature.

4.2.2) Implicit features.

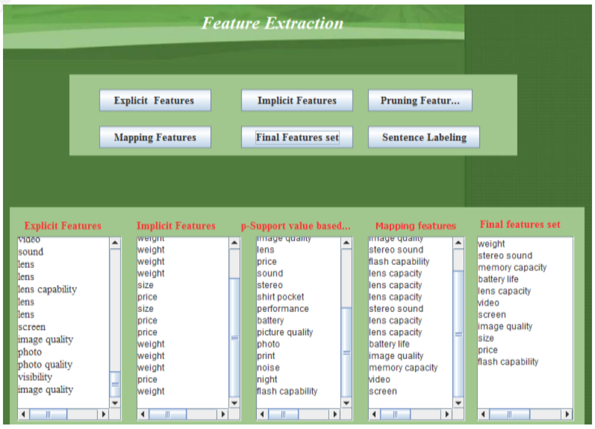
Implicit features are the product features that appear in specific adjective form in the reviews. The amounts of implicit features are few in the reviews. The adjective almost has clear meaning, so one adjective can be mapped to one noun. So a mapping database is created to obtain the implicit features.

It is too heavy to handle.
It is too expensive to buy it.

The feature 'weight' can be understood in Sentence 1, and the price also can be got in Sentence 2. So heavy is mapped to weight, and expensive to price. In order to understand more adjectives, if there are no explicit features in the opinion sentence, then this method can be used to find the implicit feature.

4.2.3) Feature Pruning.

Feature pruning aims to remove incorrect features. Hu and Liu in [8] define p-support (pure support) of feature ftr as the number of sentences that 'ftr' appears in as a noun or noun phrase, and these sentences must contain no feature phrase, which a superset of 'ftr'. One noun can be identified with repeated nouns. So a mapping database is created to obtain the feature pruning. Redundancy feature pruning focuses on removing redundant features that contain single words. So it identifies the product features automatically and it covers all the features about the products. The Implementation output for the Feature Extraction process is given in the Figure.3.



Explicit Features	Implicit Features	p-Support value based...	Mapping features	Final features set
video	weight	image quality	image quality	weight
sound	weight	lens	stereo sound	stereo sound
lens	weight	price	flash capability	memory capacity
lens	weight	sound	lens capacity	battery life
lens capability	size	stereo	lens capacity	lens capacity
lens	price	shirt pocket	lens capacity	video
lens	size	performance	stereo sound	screen
screen	price	battery	lens capacity	image quality
image quality	price	picture quality	lens capacity	size
photo	weight	photo	battery life	price
photo quality	weight	print	image quality	flash capability
visibility	weight	noise	memory capacity	
image quality	price	night	video	
	weight	flash capability	screen	

Figure.3.Implementation output of Feature Extraction.

4.3. Sentence Labeling

Customers express their opinions about products in multiple ways. Two kinds of sentences are useful while ranking products: The different types of sentences are as given below:

*A. Subjective Sentence (SS***)**: A sentence expressing direct praise or deprecation about a product.

Ex. This camera has excellent clarity.

B. Comparative Sentence (CS): A sentence which indirectly express an opinion by performing a comparison between two products.
Ex. I think the shutter speed is better than the canon sd1200.

4.3.1) Identifying Comparative Sentence
As in [11], identification of the comparative sentences be achieved through a set of keywords. It is meant for keyword comparison. KW contains 126 words, some of which are explicit ("outperform, exceed, compare, superior, etc.") and others are implicit ("prefer, choose, like, etc."). Using only this set of words to identify comparative sentences leads to a high recall but a relatively low precision. To improve the precision, the authors analyze the semantics of a sentence and its structural patterns. To identify part-of-speech tags, Stanford part-of-speech (POS) tagger [13] for English is employed. The following rules are used for identifying comparative sentences:

- Check if the sentence contains any comparative keywords in KW;
- Scanning if any predefined structural patterns are present in the sentence (as <word> as, the same as, similar to, etc.).

For example, the sentence *"I bought this camera for my daughter because she is pursuing her degree in photography."* does not show any comparative meanings or implications over other camera products. Such sentences are called subjective sentences.

4.3.2) Identifying Sentence Sentiment Orientation
In this work, it is proposed to use a simple yet powerful method by utilizing a positive word set (POS) and a negative word set (NEG) developed. A simple technique is used to identify the orientation of a sentence using these words. For the sentences containing words in the Positive word set, a positive tag is assigned and the sentences which contain the negative word set, the negative tag is assigned. This is to identify the above mentioned sentence labeling.

4.4 Ranking Products
The rank algorithm is used to find the accurate ranking. Those features involved in sentences are considered to identify the subjective and comparative sentences. And the sentence labeling score can be calculated by dividing using sentimental orientation. The ranking algorithm is applied to customer reviews from different product categories (For example: digital camera) from Amazon.com. The total number of sentences, frequency of occurrence of different product features, number of subjective and comparative sentences and their sentiment orientations are identified.

Algorithm 1 pRank Rank Products for Feature f

Require: Product Feature (f)
Ensure: The ranking list of products for the mining product feature f.

```
LSENT = Label (SENT, F);
{PS, NS, PC, NC} → LSENT;
for each sentence s € {PC, NC} do
    Find all product comparison pairs {$p_i$; $p_j$} using
    dynamic programming;
    Pair ← { $p_i$; $p_j$} + 'Pos' or 'Neg';
end for
Count $PS_{pi}$, $NS_{pi}$, $PC_{pi}$; $_{pj}$ ,$NC_{pi,\ pj}$ related to all products;
for each product $p_i$ do
    for each product $p_j$ do
        if i == j then
            Matrix [i; i] = $PSp_i/NSp_i$;
        else
            Matrix [i; j] = PC $p_j$; $p_i$/NC $p_j$; $p_i$;
        end if
    end for
end for
Ranking List = Rank (Matrix []);
return Ranking List;
```

In this above algorithm sentence labeling is first assigned. From that the count for the sentence and sentimental orientation is then ranked by dividing positive subjective and negative subjective sentences. Similarly comparative sentences are also processed. A relevant set of customer interested products features are to be ranked. It will give the accurate ranking for each feature of a particular product. Similarly, if a product has high overall rank then it should rank highly according to all mining features based on sentimental orientation.

4.5 Performance Evaluation

Performance evaluation is an important task since it determines the efficiency of the system that is developed. The evaluation can be done based on various criteria and in the proposed system well known evaluation measures like precision and recall are used.

$$\text{Precision (P)} = \frac{\text{No. of relevant document retrieved}}{\text{Total no. of document retrieved}}$$

$$\text{Recall(R)or(TP)} = \frac{\text{No. of relevant document retrieved}}{\text{Total no. of existing relevant document}}$$

Precision = tp / tp + fp
Recall = tp / tp + fn
True Negative Rate = tn / tn + fp
Accuracy = tp + tn / tp + tn + fp + fn

Note: True positive tp, False positive fp, False negative fn, True negative tn. In order to evaluate the proposed system available large datasets are to be collected.

5. RESULTS AND DISCUSSIONS

In the proposed system, the mining and ranking of product features automatically from the online customer reviews is to be done. Sentence labeling is to be done for sentimental orientation. The resulting system is expected to produce a better precision and recall compared to the existing system.

REFERENCES

[1] Alok Choudhary, Zhang K., Narayanan R., and Choudhary A., (2009) 'Mining Online Customer Reviews for Ranking Products', Technical Report, EECS department, Northwestern University.

[2] Ana-Maria Popesu and Oren Etzioni, (2005) 'Extracting Product Features and Opinions from Reviews', Proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, ACL, Vancouver, pp. 339-346.

[3] Andrea Esuli and Fabrizio Sebastiani, (2006) 'SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining', In Proceedings of LREC-06, 5th Conference on Language Resources and Evaluation, Genova, IT, pp. 417-422.

[4] Bing Liu, Mingqing Hu and Junsheng Cheng, (2005) 'Opinion Observer: Analyzing and Comparing Opinions on the Web', WWW, ACM, Chiba Japan.

[5] Etzioni O., Cafarella et al. M., (2005) 'Unsupervised named-entity extraction from the web: An experimental study', Artificial Intelligence, pp. 91-134.

[6] Kunpeng Zhang Ramanathan Narayanan, (2010) 'Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking', Electrical Engineering and Computer Science Department Northwestern University.

[7] Mingqing Hu and Bing Liu, 8 (2004), 'Mining and Summarizing Customer Reviews', Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD-2004), pp. 168–174.

[8] Mingqing Hu and Bing Liu, (2004) 'Mining Opinion Features in Customer Reviews, American Association for Artificial Intelligence.

[9] MPQAcorpus http://www.cs.pitt.edu/mpqa,2002.

[10] Mingqing Hu and Bing Liu, 7 (2006) 'Opinion Extraction and Summarization on the Web', AAAI, pp. 1621-1624.

[11] Nitin Jindal, and Bing Liu, 8 (2006), 'Identifying Comparative Sentences in Text Documents', SIGIR. pp. 244-251.

[12] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, JanyceWiebe, Yejin Choi, Claire Cardie, Ellen Riloff, Siddharth Patwardhan (2005). 'Opinion Finder: A system for subjectivity analysis', Proceedings of HLT/EMNLP 2005 Interactive Demonstrations (Demo).

[13] Weishu Hu, Zhiguo Gong, Jingzhi Guo,(2010) 'Mining Product Features from Online Reviews' Faculty of Science and Technology University of Macau , China.

[14] Xiaowen Ding, Bing Liu and Philip S. Yu, (2008) 'A Holistic Lexicon-Based Approach to Opinion Mining', Department of Computer Science University of Illinois at Chicago.