A FRAMEWORK FOR RANKING PRODUCTS USING RANKED VOTING METHOD

Rakesh Kumar
School Of Computer And System Sciences
Jawaharlal Nehru University
New Delhi, India
rakesh.kmr2509@gmail.com

Aditi Sharan
School Of Computer And System Sciences
Jawaharlal Nehru University
New Delhi, India
aditisharan@gmail.com

Payal Biswas
School Of Computer And System Sciences
Jawaharlal Nehru University
New Delhi, India
payal.biswas786@gmail.com

***Abstract*— Consumer's reviews provided with the product descriptions play a great role in the popularity of E-commerce Web sites. In order to gain confidence in the products a large number of On-line customers spend a lot of time in analyzing different textual reviews. However, there are various products, which have thousands of user's generated reviews. Mining this enormous online reviews and tuning these abundant individual consumers view into collective consumer's choice became a challenging task. These collective reviews aid in product improvement processes, ranking of various products, and many other such operations. To solve this problem, we are proposing a ranking mechanism which can be efficiently used to rank different products in accordance to their reviews rating. Here, the ranking mechanism uses the numerous ratings of a review and calculates the aggregate score of the product. This paper demonstrates that the ranking of various products by means of their reviews rating through rank voting method. Both the practicability and the benefits of the suggested approach are illustrated through an example.**

Keywords—Review Classification; Product Ranking; Rank Voting Method

I. INTRODUCTION

Web 2.0 centers around the user participation. In such scenario, sharing the opinions and sentiments about particular product or services with other people through posting online reviews has become a popular scheme. Businesses men always need to identify opinions of public regarding their respective products and services. People always wish to know the reviews of regular customers or users before buying a product. Many e-commerce websites such as Amazon (www.amazon.com) commonly provide venues sites and facilities for the users to share their reviews. Reviews are also common in social networking websites, blog posts, and many dedicated review websites such as Epinions (www.epinions.com).

A plenty of adequate information can be presented by these online reviews on various products and services and can help dealers by providing valuable network and social intelligence for the betterment of their respective businesses. Estimation of numerous available online reviews would facilitate different business person and other interested parties to avail useful information that could be economically beneficial. As a result of above scenario opinion review mining has recently gained the interest of various researchers working in the field of text analysis and opinion mining.

"Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes"[1]. It represents a wide problem space. This online word-of-mouth represents new and measurable sources of information with many practical applications.

Today, it has become a regular process among the both on-line and off-line consumers to keep them self updated about the reviews of any particular product from online web sites before going to purchase. This leads to useful customer reviews on e-commerce Web sites. Because of this, in order to seek confidence in a product, potential customers habitually brows through a large number of on-line reviews prior to purchasing. Moreover, reviews play an important and vital role to appraise the quality of products on-line. However, enormously increasing volume of reviews has led to another problem of information overload. To deal with these problems, we have proposed a product-ranking approach using reviews rating, which can help customer in choosing best products. Proposed framework for ranking products is based on rank voting method, which aims to automatically identify important products using consumer reviews rating. Our main contribution in this framework is ranking approach, which take input as reviews rating and rank the products on the bases of reviews rating.

The paper is organized as follows: Section 2 discuss the general framework for sentiment classification of reviews. in next Section 3 proposes the framework for ranking products based on Ranked voting method. Next section, we illustrate of our method with example. Finally, we conclude the paper with a summary and directions for future work in Section 5.

II. GENERAL FRAMEWORK FOR SYSTEM

This section present the general framework used for sentiment classification as shown below in Figure.1. This framework is used to classify the reviews into different classes on the basses of their strength. The various steps use in this framework are discussed below in detail.

978-1-5090-0210-8/16 $31.00 © 2016 IEEE
DOI 10.1109/CICT.2016.138

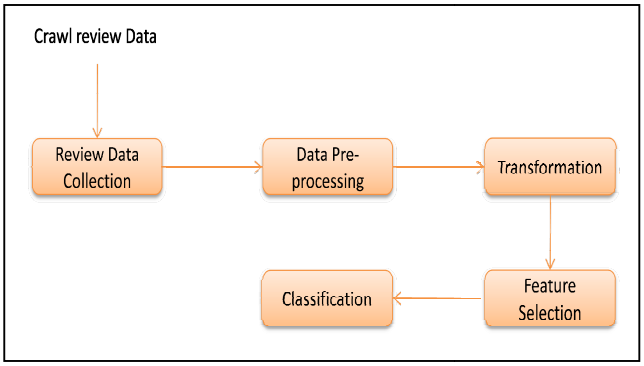

Figure 1. General Framework For Sentiment Classification[2]

A. *Review Data Collection*

Reviews are general opinion about various product, movie, books etc. given by the users. These reviews are used to find out the opinion of users about that product. There are various review dataset available online.

B. *Data pre-processing[3]*

Data pre-processing is the process of cleaning the data in order to prepare it for classification. Online raw text usually contains plenty of noise and many irrelevant things such as advertisements, HTML tags, scripts etc. In addition, whenever we go for words level processing, each word in the text is treated as one dimension. However, there are many words in the text like stopwords, which do not have a major impact on the general orientation of the text. Keeping such irrelevant words in the experimental dataset increases the dimensionality of the problem and hence makes the process of classification more difficult. It is the hypothesis that if the working data is properly pre-processed then it can reduce the noise in the text which in terns, speed up the classification process and helps in improving the performance of the classifier. The process of data pre-processing includes stop words removal like prepositions and articles, white space removal, expanding abbreviation, stemming that is to reduce term variations to a single representation etc.

C. *Transformation [4]*

In transformation, textual data is represented in numeric form. For transforming the text in numeric form binary representation is widely used. It looks for the presence or absence of a term in a document. Term Frequency (TF) that is, the number of times a term occurring in the document (i.e., term frequency) is also used as a weighting scheme for textual data. TF-IDF (Term Frequency – Inverse Document Frequency) is one of the most popular representations and considers not only term frequencies in a document, but also the relevance of a term in the entire collection of documents.

D. *Feature selection[5]*

Feature selection is a process where we run through the corpus before the classifier has been trained and remove any unnecessary features. This allows the classifier to fit a model to the problem set more quickly since there is less information to consider, and thus allows it to classify items faster. Feature selection is an important part for optimizing the performance of classifier .The main goal of the feature selection is to decrease the dimensionality of the feature space and thus improves the computational cost. Feature selection also reduces the over-fitting of the learning scheme to the training data. In Standard classification problem, we have a set of input and based upon the criteria the classifier system select appropriate features from the corpus using feature selection techniques. The features are then represented as feature vector and given as input to the classifier system.

E. *Classification[6]*

Classifier uses appropriate machine learning algorithm to classifies document into various classes. The classification could be binary or multiclass. In case of sentiment classification we can have a binary classifier predicting the negative or positive class for a classifier. On the other hand, one can also have a classifier which predicts the strength the sentiments for a review document. in such case a star rating from 1 to k (Generally k=5) is predicted to indicate the strength of sentiment.

Given the classified output for different review data, the information can be utilized to rank the product based on strength of reviews for that product. our system capture this idea to extend the functionality of existing system. The detail of proposed system presented in next section.

III. FRAMEWORK FOR PROPOSED SYSTEM

The main contribution of this paper is to suggest a product ranking mechanism based on the strength of reviews of the product. In our proposed ranking mechanism we consider similar type of products. The proposed framework is the extension of the general framework of sentiment classification as discussed in previous section. Here, the rating of the review corresponds to the class of the review. The proposed framework is a simple and effective, which ensure good product selection as well as return top-k efficient product to the customers. The input to our system is the set of review and their corresponding rating. output is top-k efferent product based on the efficiency of review. To assign a rank to the product through reviews, there are some general step that has to follow:

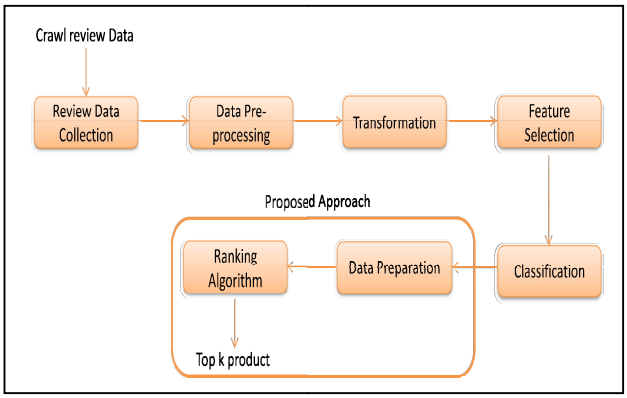

Figure 2. Proposed Framework For Product Ranking[2]

A. Data preparation for Ranking Algorithm

From the classified reviews, ranked voting data set is prepared. The format of the dataset is shown below (Table 1). Let m be the numbers of products in market and $k(k \leq m)$ be the number of numerical rating from the numbers 1 to k (here k=5). Let r_{ij} be the number of j_{th} place numerical rating of the product in i_{th} place where $i = 1 \ldots m$ and $j = 1 \ldots k$. Now our data set is prepared for ranked voting algorithm. In next step, we can apply ranked voting algorithm.

TABLE I. DATA REPRESENTATION

Product	$rating_1$	$rating_2$	$rating_3$	.	$rating_k$
$prod_1$	r_{11}	r_{12}	r_{13}		r_{1k}
$prod_2$	r_{21}	r_{22}	r_{23}		r_{2k}
$prod_3$	r_{31}	r_{32}	r_{33}		r_{3k}
.					
$prod_m$	r_{m1}	r_{m2}	r_{m3}		r_{mk}

B. Ranking of Products

After data preparation, the ranked voting algorithm can be applied to rank the products.

In this work, to find a best product for a user, ranked voting method is used [7]. In ranked voting system, voter ranks alternatives in order of preference. In our case the review rating corresponds to the order of preference. There is a long list of reviews to find an efficient product. Each review will act as a voter, product are candidates for them. Thus, a ranked voting data set is prepared. In research, some method has been proposed to analyze ranked voting data such as Data Envelopment Analysis (DEA) introduced by Cook and Kress [8]. But DEA often suggests more than one efficient candidate. Some methods are proposed to discriminate these efficient candidates. But order of preference may be changed because of existence of an inefficient candidate. Tsuneshi Obata and Hiroaki Ishii introduced [7] a novel method which does not use information of inefficient candidate to discriminate efficient candidates given by DEA. Proposed work considers the same method to find a best product for a user.

The ranked voting method is applied in two steps:-

a) Find efficient products: Let m be the numbers of products in market and $k(k \leq m)$ be the number of numerical rating i.e. a user has to select one product and assign a numerical rating to the product from the numbers 1 to k. Let r_{ij} be the number of j_th place numerical rating of the product i_{th} where $i = 1 \ldots m$ and $j = 1 \ldots k$. Now preference score z_i should be calculated for each product i as a weighted sum of numerical ratings with certain weight w_i, i.e.

$$z_i = \sum_{j=1}^{k} w_j r_{ij} \quad (1)$$

By using data envelopment analysis (DEA), Cook and Kress [8] have proposed a method for estimating preference scores without imposing any fixed weights from outset. Each candidates score is calculated with their most favorable weights. Their formulation is the following:

$$Z_o^* = maximize \sum_{j=1}^{k} w_j r_{ij} \quad (2)$$

subject to

$$\sum_{j=1}^{k} w_j r_{ij} \leq 1, \ i = 1, \ldots, m, \quad (3)$$

$$w_{j+1} - w_j \geq d(j, \in), \ j = 1, \ldots, k-1, \quad (4)$$

$$w_k \geq d(k, \in), \quad (5)$$

where $d(\cdot, \in)$, called the discrimination intensity function, is nonnegative and nondecreasing in $\in$, and satisfies $d(\cdot, \in) = 0$. Parameter $\in$ is nonnegative.

After applying DEA, value of z_i will be 1 for all efficient products. After the problems are solved for all products, several (not only one) products often achieve the maximum attainable score 1. We call these products efficient products. We can judge that the set of efficient products is the top group of products, but cannot single out only one best among them.

b) Discriminate efficient products : Let $\hat{z}_o$ be normalized preference score of efficient products $(z_i = 1)$ that has to be calculated. Model for ranked voting method with discrimination of efficient products is as follows.

$$1/Z_o^* = minimize \|w\|, \quad (6)$$

subject to

$$\sum_{j=1}^{k} w_j r_{oj} = 1, \quad (7)$$

$$\sum_{j=1}^{k} w_j r_{ij} \leq 1, i \neq o \quad (8)$$

$$w_{j+!} - w_j \geq d(k, \in), j = 1, \ldots, k-1, \quad (9)$$

$$w_k \geq d(k, \in), \quad (10)$$

where $d(\cdot, \in)$ called discrimination intensity function is non-negative and non-decreasing in $\in$ $\in \geq 0$ and satisfies $d(\cdot, 0) = 0$. Constraint (7) is for efficient products, constraint (8) is for products which are not efficient and constraint (9) means review of higher place may have greater importance than that of the lower place.

The normalized preference score Z_o^* is obtained as a reciprocal of the optimal value. Product with highest normalized preference score will be winner. i.e. best Product for user.

Our method does not use any information about inefficient products and the problem of changing the order of efficient products does not occur. Because there is no existence of an

inefficient product. In the next section, we present our method with an example.

IV. ILLUSTRATION OF OUR METHOD WITH EXAMPLE

TABLE II. SAMPLE DATA (M = 10, K =5)

Product	Five star	Four star	Three star	Two star	One star
Prod1	35	10	8	12	6
Prod2	20	40	30	15	8
Prod3	26	15	10	8	12
Prod4	12	10	15	20	24
Prod5	15	8	12	6	7
Prod6	33	24	10	15	8
Prod7	10	15	18	20	30
Prod8	8	12	15	25	28
Prod9	30	25	20	15	12
Prod10	25	30	30	20	15

We illustrate our method with an example (Table II). Preference scores for each product are calculated by Cook and Kresss basic model (2)–(5). Here, we use $d(\cdot, \epsilon) = 0$. Their scores are as follows:

TABLE III. PRODUCT WITH PREFERENCE SCORE

Product	Preference score
prod10	1
Prod1	1
Prod2	1
Prod6	1
Prod9	1
Prod3	0.788288
Prod7	0.774999
Prod8	0.733333
Prod4	0.674999
Prod5	0.486486

After the problems are solved for all products, several (not only one) products often achieve the maximum attainable score 1. We call these products efficient products. Hence in above shown example the products prod10, prod1, prod2, prod6 and prod9 seem to be efficient. We can judge that the set of efficient products is the top group of products, but cannot single out only one best among them.

TABLE IV. SAMPLE DATA (M = 5, K =5)

Product	Five star	Four star	Three star	Two star	One star
Prod10	25	30	30	20	15
Prod1	35	10	8	12	6
Prod2	20	40	30	15	8
Prod6	33	24	10	15	8
Prod9	30	25	20	15	12

In order to discriminate efficient products, we can apply our approach. which does not use information of inefficient product. So, we can apply our proposed approach to discriminate efficient product (Table IV) to find a best product for a user. Their scores are as follows:

TABLE V. PRODUCT WITH SCORE

Product	Score
prod1	34.99991
Prod6	31.874963
Prod2	29.99998
Prod10	27.499994
Prod9	27.352931

From the above result as shown in Table V, it is observe that the product prod1 : 34.99991 is best among them.

V. CONCLUSIONS AND FUTURE WORKS

Due to the enormous increase in online reviews there are various products which have thousands of user's generated reviews. Mining this enormous online reviews and tuning these abundant individual consumers view into collective consumer's choice became a challenging task. To solve this problem we have proposed a ranking mechanism which can be efficiently used to rank different products in accordance to their reviews rating. The rank voting method has been used to rank the products. The effectiveness of our approach has been shown through an example. Ranked voting method does not use inefficient product's information to discriminate efficient products therefore order of efficient product never changes if inefficient products are added or removed.

In future, this framework may also be applied for recommendation of the top-k products on the basis of product reviews rating. Further this approach can be apply in different application in opinion mining. In this work, our framework is only illustrated through example. Further we can extend this model on various review available in real life.

REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining", Synthesis Lectures on Human Language Technologies 5.1, pp. 1-167, 2012.

[2] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN", Expert Systems with Applications, pp. 621-633, 2012.

[3] E. Haddi, X. Liu, and Y. Shi, " The role of text pre-processing in sentiment analysis ", Procedia Computer Science, 17 , pp. 26–32, 2013.

[4] B. Liu,Web Data Mining :Exploring Hyperlinks, Contents and Usage Data , Chicago, USA: Springer-Verlag Berlin Heidelberg 2007.

[5] O. Kummer and J. Savoy. "Feature Selection in Sentiment Analysis", CORIA, Bordeaux, pp. 273–284, 2012.

[6] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[7] T. Obata, H. Ishii, “A method for discriminating efficient candidates with ranked voting,” European Journal of Operational Research, vol. 151, 2003, pp.233–237.

[8] W. D. Cook, and M. Kress, “A Data Envelopment Model for Aggregating Preference Rankings,” Management Science, Vol. 36, No. 11, pp. 1302–1310, 1990.