International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

An Improvised Ideology based K-Means Clustering Approach for Classification of Customer Reviews

Kinnari Mishra

Parul University, Vadodara, Gujarat, India
Email: *kinnari.mishra12751[at]paruluniversity.ac.in*

Abstract: ***Background/Objectives: To provide a framework for improving the classification of customer reviews on products. Methods/Statistical Analysis: We propose an integrated framework for classifying the customer reviews based on the textual analysis with constraint-based association rules using ontology. It involves preprocessing the customer reviews including symbols and handling feature extraction. An improved K-Means algorithm with ontology is proposed to consolidate the reviews based on textual analysis method to handle reviews that represent at least one feature of the product. Findings: The empirical results reveal that the accuracy of the system increases with the use of ontology and modified K-Means algorithm, improving overall performance of the recommendation system. Combining preprocessing and ontology considerably improves the accuracy of classification of customer reviews. Applications/Improvements: The proposed approach can be used to recommend product based on users' review.***

Keywords: Classification, K-Means Clustering, Ontology, Preprocessing, Recommendations, Review

1. Introduction

In the knowledge society, the products are purchased by customers based on the previous reviews of the products given by users' through social networks such as Facebook, Twitter and other forums. For instance, Amazon[1] provides a recommendation about their products and about 20% of the sales in Amazon are triggered by the reviews on the recommendation systems. However, the users' may need independent recommendation systems. The real challenge in categorizing the reviews based on the users' point of view is to analyze the sentiments conveyed in the reviews. Sentiment analysis is carried out with the sentence used by the reviewer to propose their views. Also, symbols like "smiley" play an important role in sentimental classification system.

In2 proposed a News recommendation technique utilizing real-time Twitter data as the basis for ranking and recommending articles from a collection of simple syndication feeds. It is found that the users' with more friends tend to benefit more. In3 explored three separate dimensions in designing a recommender: content sources, topic interest models for users', and social voting. They have demonstrated that both topic relevance and the social voting process were helpful in providing recommendations.

As the number of reviews given by users' is large and they have only few sentences containing opinions on the product it becomes hard for a potential customer to read them to make an informed decision on the purchase of product. Another study4 suggests mining the product features commented by customers in their review and then identifies the opinion of each review.

Thus, we are in need of a system which automatically provides the recommendation from all available resources. Automated systems are meant to automate analysis, summarize the given reviews and classify them according to the sentimental view of the text. Text mining is an inter-disciplinary method used in different fields like machine learning, information retrieval, statistics, and computational linguistics. Web mining is a sub-discipline of text mining used to mine the semi-structured web data in the form of Web Content Mining, Web Structure Mining and Web Usage mining. Opinion mining also called senti- ment analysis which is a process of finding user's opinion about a particular topic or a product or problem. For a product manufacturer, there are additional difficulties because many merchant sites may sell its products, and the manufacturer may produce many kinds of products. In[5] proposed a work on opinion mining using Machine Learning techniques for text classification and gives an overview on linguistic resources required for sentiment analysis. A method is proposed[6] to mine attributes from book reviews by identifying book features mentioned in the review text. This process identifies a set of coordinates for every book and user and measures the values corre- sponding to the global term frequency-inverse document frequency (tf-idf) for each of the feature tag words. Some other research studies have developed visualization tech- niques to assist with the identification and evaluation of keywords, patterns and emotive categories[7].

In[8] authors proposed a system for predicting the users' interests based on the relevant feedback system. In this paper, fuzzy classification is used to classify the user's inter-applied Heuristic search-enhanced Markov blanket model together with SVM for opinion mining from online text.

An enhanced K-Means algorithm[16] was conveyed to evaluate the newly evolved user groups. It also shows the group formation of users and the way they are grouped among themselves. Another work proposed in[17] discussed about the summarization of online reviews based on user preferences. In this work, the author explains about the relevance feedback mechanisms that are implicitly given by the users during their purchase. The same kind of work is discussed in[18] to show the ranking of reviews based on the comments of users. They rank the user reviews by using the weight factor of the words in the user reviews. Thus, the reviews are clustered not only based on the sentence formation but also based on the sentimental anal- ysis of the words usage. Here, in section 2 we discuss about Amazon dataset. Section 3 deals with the

architecture of the clustering system. Section 4 explains the implementation details. Finally, section 5 depicts the results and evaluation of the proposed system and shows the better clustering way to improve the recommendation systems.

2. Dataset

It also explains the approach for ranking the users' interest. Human input however plays an important role in sentimental analysis. This can include approaches that rely on human interaction as part of the initial identification of feature from content, as well as methods that use human interaction as a tool of evaluating the results of algorithm- based methods to produce these results[9,10]. Techniques to summarize and categorize data are still largely dependent on human evaluation to generate meaningful results and will likely remain so for the foreseeable future.

In[11] used Naive Bayes (NB) and reported that NB gave the best result to restaurant reviews and obtained 83.6% accuracy on more than 6000 documents. The approach in[12] was based on lexical WordNet together with NB, Support Vector Machines and decision tree for sentiment classification. They reported a performance accuracy of greater than 75%.

Sentiment classification approach based on latent semantic analysis (LSA) to identify product features was adopted by[13] and proposed a system called a movie-rating and review-summarization system in a mobile environment. In[14] proposed TF-IDF weighting schemes combined with SVM classifier. This solution achieved a significant improvement over the previous study. In[15] proposed an Amazon is the one of the most popular commercial network sites all over the world. It sells books, music, electronics, household items and other human needed things. The customers review on products and the overall ratings given by them are published on the product page. We crawl the ratings given by the reviewers and link the words with the review and group them according to 5 different categories. Users in this dataset have social relationships based on the products they buy in the sites. We have crawled nearly thousand users' circles of friends and their ratings from December 2010 to March 2015. We first collected review of active users in Amazon, who give more than five reviews for different books and also check the origin of the user. We then further crawl these users' friends to build region-wise sub-networks of Amazon. Except the user without rated history (at least one rated item) and friends (at least one friend), the dataset consists of ratings from users who rated a total of nearly 1,00,000 books.

3. System Architecture

The architecture of the proposed system is shown in Figure 1. The reviews from the data set are crawled and stored in the database. We store the review data based

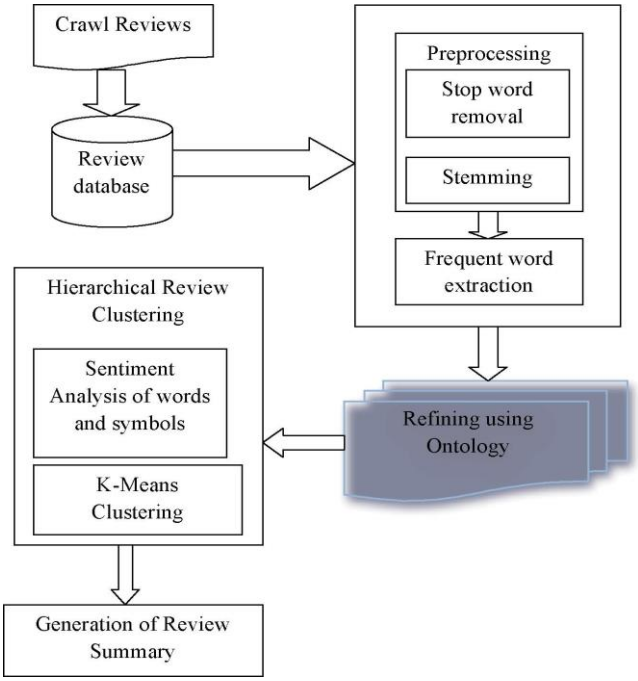

Figure 1: Overall System Architecture

on the products and features of products in the database. The operations on data start from the preprocessing step which includes stop word removal and stemming. Stop word removal is the linguistic process used to remove words such as prepositions, which reduces the operating space for processing the reviews. After carrying out stop word removal we improve our preprocessing to identify review features and the other symbols used in the reviews as some reviews may have smiley's like ':)' and in some it would be separated by ':' and ')' which is not needed. But when the symbols appear together it can be taken as a smiley. To implement this we make our system learn about the types of symbols. The next preprocessing step is stemming which removes the suffix like 'ing', 'ful', 'ness' etc., from the root word. But we consider the suffixes for feature review generation. For example, a stem word 'friend' could be mentioned as 'friendly' and somewhere the same word could be mentioned as 'friendliness'. If only the stem word is taken after stemming it provides a positive kind of view for both the cases. We use the stemming process to understand the meaning of the root word. After stemming, we extract the frequent word from the reviews. All the reviews will have some common words like good, beautiful, etc., which we group into positive and negative reviews.

Ontology generation is used to generate a concept map- ping between the review contexts. In general, reviews are categorized as positive, negative and neutral. But the words which related to that particular category are mapped using ontology in a hierarchical manner. For example, 'informative' is a word in the review which gives the positive thought mapped under the positive review. The ontology improves the accuracy by refining the attributes.

We use the hierarchical review clustering which comprises of two components, sentiment analysis of word or symbols and K-Means Clustering algorithm. In this sentiment analysis, rule generation and classification are done. We classify the features of the products based on sentimental words in used

in the text and also, we measure the level of reviewers based on the quality of their past reviews. The context between their reviews is com- pared with the use of ontology mapping concepts. Review context is classified by using the concept mapping of ontology. We use modified K-Means algorithm to classify the review with ontology mapping.

A review context *R* contains a set of attributes which we consider as classes root word and is only for understanding the base meaning of the comment but not evaluating the reviews of the user. Then, we perform the frequent word extraction to retrieve the frequent occurrence words in the reviewers comment. Here, Euclidean distance metrics is used to evaluate the similarity between the words. When compared to tf-idf, Euclidean can able to find the similar sentences in the overall comments of the reviews. It also helps to find the similarity in multi-dimensional words in bag of words.

In general, the equation which is used to find the distance between two vectors in n-dimensional space is *D*

$$D2(a, b) = (a_1 - b_1)^2 + (a_2 - b_2)^2 + ...+(a_n - b_n)^2 \quad (1)$$

The modified K-Means algorithm is summarized as:

1) The review context set, $S = \{s^1, s^2, s^n\}$, where every s^i is considered as an attribute.

If s^i belongs to the correct class

$$s^i \in S$$

else

s^i is an outlier delete s^i from *S*

2) Create a new database, $D = \{p^1, p^2, ...p^m\}$, where every p^i is considered as a point

3) For every point p^i from the original set, identify the s^i based on points in new database
   - If p^i is an absorbed point, then retain p^i in *D*
   - If not, then remove p^i from *D* and add it new database of prototypes

4) Proceed with all elements in the original set.

5) Repeat steps 3 and 4 until no new concepts are added.

Once the classification is carried out for both symbols and context, the summarization is done based on the classified data. The improved algorithm provides better results as the Section 5 reveals it.

Implementation and Results

As a case study to illustrate our proposal the data set.

crawled from Amazon is taken as a review database for our experiments. It consists of about 100 thousand but the

quality of the reviews is not satisfactory. We filtered the reviews and extracted 15,000 complete reviews for the books. The reviews are given as input to the preprocessing stage. In preprocessing, stop word removal and stemming process is carried out. Null stemmer is used for stemming the words and processed without the stop word removal. Here, most of the words are adjective types because the reviews mostly describe the content of the product. With respect to books the reviewers comment more on the book quality by the way it describes the subject and the way it is organized. As already discussed, stemming is to find the n □ i i i□1

K- Means Clustering

The hierarchical structure of the review comments is generated from the protege tool. Weka tool is used to cluster the data based on the sentence similarity of the comments. All the reviews (12,75,725) given to our system to cluster them into groups. First, the instances are created for the given reviews using ontology. From nearly ten lakh reviews the 2000 instances are created. From these 1353 (67.65 %) instances that are helpful for clustering the reviews are extracted and made as a label. The accuracy of the clustered instances are verified by the following methods

- Mean Absolute Error (MAE),
- Root Mean Squared Error (RMSE)
- Relative Absolute Error (RAE).

Mean Absolute Error (MAE)

MAE is the average of the absolute errors, where f^i is the prediction and y^i is the true value. For our instance of data we have identified that the mean absolute error rate is 0.32.

$$MAE = MAE = \frac{1}{n}\sum_{i=1}^{n} |f^i - y^i \vee| \quad (2)$$

Root Mean Squared Error (RMSE)

RMSE is the average of the square of all the error. RMSE identifies a large number of errors compared to MAE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a - a)^2} \quad (3)$$

RMSE has improved the identification of error rate when compared to MAE. Here, the identified error rate 0.5535 whereas in MAE is 0.32.

3. Result Analysis

We cluster the instances according to the features that appear in the reviews. Reviews are generally taken as sen- tences that are context attached to a single word which gives the meaning of entire review. For instance, the review for book may be “book quality is good, more informative, lots of examples”. In this review “book quality is good” gives the same meaning as “book is good”. We get the same meaning but the first one insists the book quality and second says about book in general. Here, the quality word cannot always act as a positive review because if we mention bad instead of good then it becomes a bad review. To handle such contradicting issues we create instances for the reviews using ontology that helps in classifying the reviews properly.

We carried out the following analysis

- Effectiveness of preprocessing.
- Effectiveness of instance creation using ontology.

- Effectiveness in true positive rate and false positive rate.
- The accuracy of classification based on the reviews.

The results obtained for our case study is given in Table 1, Table 2, Table 3 and Table respectively.

Table 1: Result analysis for Positive and Negative reviews without performing preprocessing and ontology

Class	True Positive	False Positive	Precision	Recall	F-measure
Positive	0.479	0.663	0.419	0.479	0.447
Negative	0.337	0.521	0.392	0.337	0.545

Table 2: Result analysis for Positive and Negative reviews after performing preprocessing steps and without ontology

Class	True Positive	False Positive	Precision	Recall	F-measure
Positive	0.568	0.516	0.523	0.568	0.618
Negative	0.484	0.432	0.528	0.484	0.363

When we compare Table 1 and Table 2, F-measure, Precision and Recall rates are significantly higher in Table 2 than Table 1. It is because of performing preprocessing steps for the crawled dataset.

Table 3: Result analysis for Positive and Negative reviews with ontology instances without performing preprocessing steps

Class	True Positive	False Positive	Precision	Recall	F-measure
Positive	0.686	0.535	0.561	0.686	0.505
Negative	0.465	0.314	0.596	0.465	0.523

Table 4: Result analysis for Positive and Negative reviews with preprocessing steps and ontology instances

Class	True Positive	False Positive	Precision	Recall	F-measure
Positive	0.814	0.461	0.638	0.814	0.716
Negative	0.539	0.186	0.743	0.539	0.625

The results show that the performance of the system is better when we use preprocessing and ontology concept generation together. The results reveal that the true positive rate increases in prediction with ontology and preprocessing steps. The other performance features also shows a significant increase leading to better prediction.

4. Conclusions

The developments in Information and Communication Technology (ICT) have increased the growth of customer reviews on products due to sheer volume. It remains a challenging task to classify the reviews given by customers for same or different products. In this paper, we propose a new integrated framework by which we consolidate the reviews given by different users based on the features the buyer wants. The approach also provides a view for manufacturers in predicting the future expectations of their customers. Textual analysis techniques are used to recognize the symbols, stickers, and smileys used by the user while expressing their reviews. We deploy ontology along with modified K-Means clustering in classifying the reviews that are understandable both by customers and manufacturers. It is observed from the experimented results that the proposed approach of combining ontology with K-Means clustering provides improved performance for classifying Customer Reviews.

References

[1] Linden Greg, Smith Brent, York Jeremy. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing. 2003; 7(1):76–80.
[2] Phelan O, McCarthy, myth B. Using twitter to recommend real time topical news. Hong Kong, China: Proc. 3rd ACM Conf. Recommender Systems. 2009; p. 385–88.
[3] /Chen J, Nairn R, Nelson L, Bernstein M, Chi E. Short and tweet: Experiments on recommending content from information streams. Atlanta, GA, USA: Proc. 28th Int. Conf. Human Factors Computing Systems. 2010; p. 1185–94.
[4] Minqing Hu, Bing Liu. Seattle, Washington, USA: Mining and Summarizing Customer Reviews KDD'04. 2004 August 22–25.
[5] Padmaja S et al. Opinion Mining and Sentiment Analysis - An Assessment of Peoples' Belief: A Survey. International Journal of Ad hoc, Sensor & Ubiquitous Computing IJASUC. 2013 Feb; 4(1).
[6] Lin Eric, Fang Shiaofen, Wang Jie. Mining Online Book Reviews for Sentimental Clustering. 27th International Conference on Advanced Information Networking and Applications Workshops. 2013; p. 179–84.
[7] Oelke D, Bak P, Keim D, Last M, Danon G. Visual evaluation of text features for document summarization and analysis. IEEE Symposium on Visual Analytics and Technology. 2008; p. 75–82.
[8] Sai Ramesh L, Ganapthy S, Bhuvaneswari R, Kannan A, Kulothungan K, Pandiyaraju V. Prediction of User Interests for Providing Relevant Information Using Relevance Feedback and Re-ranking. International Journal of Intelligent Information Technologies. 2015; 11(4):1–17.
[9] Goldberg D, Nichols D, Oki BM, and Terry D. Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM. 1992; 35(12):51–60.
[10] Hill W, Stead L, Rosenstein M, and Furnas G. Recommending and evaluating choices in virtual community of use. Proceedings of CHI'95. 1995; p. 194–201.
[11] Kang H, Yoo SJ, Han D. Senti-lexicon and improved. Naive Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications. 2011.
[12] Annett M, Kondrak G. A comparison of sentiment analysis techniques: Polarizing movie blogs. Advances in Artificial Intelligence. 2008; 5032:25-35.
[13] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou. Movie Rating and Review Summarization in Mobile Environment. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews. 2012 May; 42(3).
[14] Paltoglou G, Thelwall M. A study of information retrieval weighting schemes for sentiment analysis. Proceedings of the 48th Annual Meeting of the

Association for Computational Linguistics. 2010; p. 1386–395.

[15] Rui Xia, Chengqing Zong, Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment clas- sification. Information Sciences. 2011; 181:1138-52.

[16] Selvakumar K, Sai Ramesh L and Kannan A. Enhanced K-Means Clustering Algorithm for evolving User Groups. Indian Journal of Science and Technology. 2015 Sep; 8(24). Doi: 10.17485/ijst/2015/v8i24/80192.

[17] Prakash S, Chakravarthy TC, Brindha GR. Preference Based Quantified Summarization of On-line Review. Indian Journal of Science and Technology. 2014 Jan; 7(11). Doi: 10.17485/ijst/2014/v7i11/51935.

[18] Meghana Ramya Shri J, Subramaniyaswamy V. An Effective Approach to Rank Reviews Based on Relevance by Weighting Method. Indian Journal of Science and Technology. 2015 June; 8(11). Doi: 10.17485/ijst/2015/ v8i11/61768