RESEARCH ARTICLE OPEN ACCESS

Multi Document Utility Presentation Using Sentiment Analysis

Mayur S. Dhote [1], Prof. S. S. Sonawane [2]
Department of Computer Science and Engineering
PICT, Savitribai Phule Pune University
Pune - India

ABSTRACT

Today's web is packed with billions of pages with information on any topic. While finding information on the specific topic, there are chances of skipping important pages with important information. This may create negative impact toward topic, person, and organization. To resolve this problem, solution is propose in this paper that is, "Multi Document Utility Presentation using Sentiment Analysis". In propose method, relevant information available on the web is associated with keyword which can be direct, indirect or subordinate and based on this sentimental analysis is performed on the data. This will help the user for decision making. To identify indirect reviews, concept of keyword extraction and semantic similarity is used. The results obtained show that consideration of indirect reviews gives impact on rating of each aspect. The cluster of each aspect is formed by combining BON (Bag of Noun) and semantic similarity methods. The goodness of clusters is measured by purity method as shown in the result. This system helps user to understand polarity of each aspect of product while purchasing.

Keywords:- Clustering, Sentiment analysis, Text Mining, User Reviews.

I. INTRODUCTION

The process of analyzing people's reviews or opinion about different living or non-living things is known as sentiment analysis or opinion mining. The sentiment analysis becoming a most growing research area as the organization brings the use of e-commerce into the operation. The most of research in text mining is focused on extraction or summarization of textual information. But the main task of sentiment analysis is to extract opinions of the peoples from the textual information available on the web.

The web user and organization both can take advantage of opinion mining. When an individual wants to buy a product he\she may refer reviews given by people on different website. But reading and analyzing those reviews could be lengthy and perhaps frustrating process. Organization to improve their services there is need for analyzing review. With sentiment analysis is we can automatically analyze large amount of data and extract opinion.

Until now most researchers have considered only direct review while making a decision or analyzing sentiments. But due to increase in web data it is also necessary to consider indirect or subordinates review in sentiment analysis process to increase accuracy.

There are three different levels of sentiment analysis, 1) Document level, 2) Sentence level, 3) Aspect level. In the document level sentiment of whole document should be considered. In sentence level the sentiment of each sentence in the document is considered. And in aspect level sentiment of each aspect is considered rather than considering the whole product sentiment.

In this paper, aspect level sentiment analysis use by considering both direct and indirect review. After getting all related review next step is to identify aspects. Aspects also called as features of an object. Example, if we consider 'Mobile' then 'screen', 'audio', 'camera', etc are the aspects or features for a mobile. To find aspects follow idea of employing clustering over the sentences. At last calculate polarity of each aspect to identify rating.

The remaining part of the paper describes as follows, section II describes the related work. Section III defines the proposed system. Section IV shows the result generated and section V concludes the paper.

II. RELATED WORK

Michael Gaman et. al[1] designed a prototype system used for a car review database. The proposed system combined the clustering and sentiment analysis process together to produce an effective result.

Chunliang Zhang et. al[2] proposed an unsupervised techniques for classification as supervised technique is an expensive. The multi-class bootstrapping algorithm used to remove ambiguity i.e. single term related to more than one aspect of aspect related term. For identification of aspect, it uses a single class bootstrapping algorithm and multi-class bootstrapping aspect related degree value.

Jingbo Zhu et. al[3] proposed a multi-aspect bootstrapping method to identify a term related to aspect. Author also proposed an aspect based segmentation model to partition a sentences which contain multiple aspects into single aspect sentence.

Mostafa Koramibakr et. al[4] justify the importance of the verb when we are considering reviews on social issues. So for this purposed authors create an opinion verb dictionary which defines a polarity of verb and also created a opinion structure which help when there is more than one opinion word in different review related to social issues.

Xu Xueke et. al[5] proposed Joint aspect/sentiment model which uses both LDA (Latent Dirichlet Allocation) and JAS (Joint Aspect Sentiment) model. Using LDA model extracts all the topic related to the aspect. After extraction of topics it uses JAS model to extract the aspects and aspect dependant terms.

Raisa Varghese et. al[6] proposed a sentiment analysis model. In these model reviews are classified according to it's useful or not using sentiwordnet. The aspect should be identified using proper POS tagging by identifying noun and pronoun. The co-reference resolution is resolve using Standford Deterministic Coreference Resolution system.

From the above related work it is observed that consideration of only direct review not make the system effective. So there is a need to consider indirect review make system effective. It is also observed that forming a cluster[7],[8],[9] of the review to find aspect is better approach than analyzing each review separately.

III. THE PROPOSED TECHNIQUES

Figure 1 gives the architectural overview of the utility identification system. The user query is provided as input to the system. After getting an input review should be crawled from review database to get direct and indirect reviews. While getting direct reviews we have to split sentences of review dataset to calculate the similarity between sentence and user query. Indirect reviews extracted by finding a semantic similarity between web information extracted using keywords and user query. After getting a reviews BON approach is used to identify aspects. A cluster of reviews should be formed using both BON and semantic approaches. Opinion word extracted from the each cluster to find the polarity of each aspect. At last the rating evaluation should be done on the basis of the polarity of aspects.

A. MODULE DESCRIPTION

1) Pre-processing:

The pre-processing includes sentence splitting and tokenization. For sentence splitting 'en-sent.bin', a pre-trained model is used. Consider the following example of sentence splitting as shown in figure 2.

2) Direct Reviews Identification:

To identify direct reviews cosine similarity method is use. The cosine similarity between the user query and each review in review dataset is calculated. If similarity value is greater than the threshold value then the review is selected as a direct review.

- Cosine Similarity[10][11]:

The cosine similarity is a technique to identify the relatedness between the sentences or documents. The cosine similarity should be calculated using the following formula:

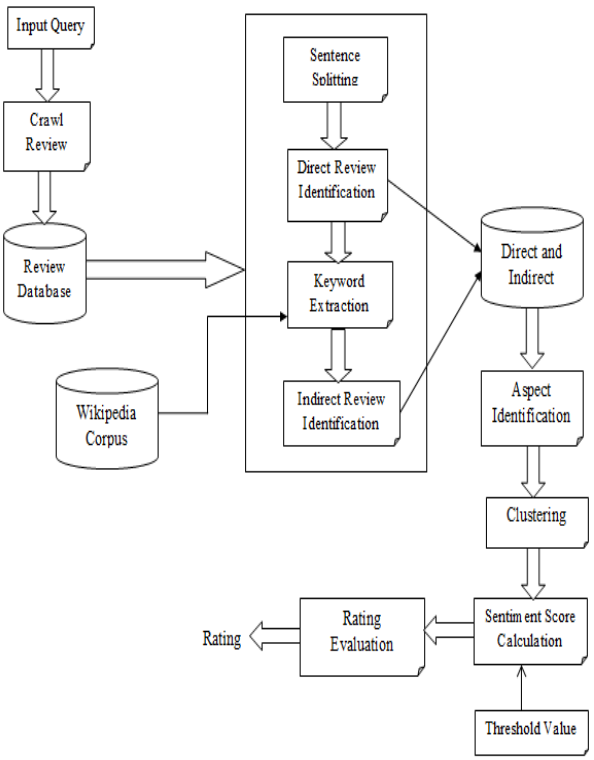

Fig. 1 Architectural Overview

The apple provides best smartphone in market. Sound of Samsung phone is not good. Google launches smartphone.

Sentence Splitting

The apple provides best smartphone in market.

Sound of Samsung phone is not good.

Google launches smartphone.

Fig. 2 An example of sentence splitting

$$\text{Cosine}(Q, r_i) = \frac{Q \cdot ri}{||Q|| \, ||ri||} \ldots..(1)$$

$$\text{Cosine}(Q, r_i) = \frac{\sum_{i=1}^{n} V_{Qi} V_{ri}}{\sqrt{\sum_{i=1}^{n} V_{Qi}^2} \sqrt{\sum_{i=1}^{n} V_{ri}^2}} \ldots..(2)$$

Where,

Q = User query

r_i = Review from review database.

VQi = Vector space of words in user query.

Vri = Vector space of words in reviews.

The benefits of using cosine similarity is,

- It provides model based on the simple linear algebra
- Terms are represent using weight not by binary value

Algorithm 1: Identify_Direct_Reviews

Input: The user query 'Q', set of reviews R = $\{r_1, r_2, \ldots, r_n\}$ and threshold 'th'

Output: A set of direct review R' = $\{r_1, r_2, \ldots, r_n\}$.

Procedure:

1: For each sentence S ϵ R and query 'Q'
2: Derive vector i.e. V = {v1,v2,…,vn}.
3: Calculate Cosine Similarity between sentence and query
C = Cosine(S,Q)
4: If (C >th) /* 'th' is threshold value */
5: Extract sentence and stored into text file
DR = $\{S_1, S_2, \ldots, S_n\}$ /* Set of Direct Reviews*/
6: END If
7: END For

---

3) Keywords Extraction:

To identifying keywords, TF-IDF method is used. The keywords which are related to the domain of user query are extracted. The TF-IDF can be measure using following formulas[12][13],

$$\text{Term Frequency (TF)} = \frac{Freq(T)}{Num(T)} \ldots..(3)$$

$$\text{Inverse Document Frequency (IDF)} = \frac{Num(D)}{Doc\_Freq(T)} \ldots..(4)$$

$$\text{TF-IDF} = \text{TF} * \text{IDF} \ldots\ldots..(5)$$

4) Indirect Review Identification:

To identify indirect reviews following steps need to follow:

1) Extract URL related to every keyword extracted from the previous module. To extract URL, Google API is used.

2) After getting a URL, the data containing on that URL was extracted and stored into another text file.

Semantic similarity is calculated between user query and web pages data to identify indirect reviews. To calculate Semantic Similarity following steps need to be follow[14]:

Step 1: Words Similarity

This step determines the value of Noun Vector (NV) and Verb Vector (VV) for each sentence.

$$NV_{Si} = \sum_{k=1}^{|S\_N_{Si} \cup S\_N_Q|} MAX(Similarity(WORD\_Si, NOUN\_BASE_k))$$

$$VV_{Si} = \sum_{k=1}^{|S\_V_{Si} \cup S\_V_Q|} MAX(Similarity(WORD\_Si, VERB\_BASE_k))$$

Where,

S_N_{Si} = Set of noun in sentence 'S_i'

S_N_Q = Set of noun in query 'Q'

S_V_{Si} = Set of verb in sentence 'S_i'

S_V_Q = Set of verb in query 'Q'

Similarity(Words_S_i, Noun_Base$_k$) = 2 * Depth(H1) * Depth_length(Word_S_i, H1) * Depth_length(Noun_Base$_k$) + 2*Depth(H1)$^{-1}$

Where,

H1 = Depth of lowest shared hyponym of Word_S_i and Noun_Base$_k$

Step 2: Cosine Measurement

$$\text{Noun Cosine} = NC_{Si,Q} = \left(\frac{\overrightarrow{NV}_{SEN\_Si} \cdot \overrightarrow{NV}_{SEN\_Q}}{|\overrightarrow{NV}_{SEN\_Si}| * |\overrightarrow{NV}_{SEN\_Q}|}\right)^2$$

$$= \left(\frac{\sum_{i=1}^{|S\_V_A \cup S\_V_B|} NV_{SEN\_A_i} \cdot NV_{SEN\_B_i}}{\sqrt{NV^2_{SEN\_A_i}} * \sqrt{NV^2_{SEN\_B_i}}}\right)^2$$

$$\text{Verb Cosine} = VC_{Si,Q} = \left(\frac{\overrightarrow{VV}_{SEN\_Si} \cdot \overrightarrow{VV}_{SEN\_Q}}{|\overrightarrow{VV}_{SEN\_Si}| * |\overrightarrow{VV}_{SEN\_Q}|}\right)^2$$

$$= \left(\frac{\sum_{i=1}^{|S\_V_A \cup S\_V_B|} VV_{SEN\_A_i} \cdot VV_{SEN\_B_i}}{\sqrt{VV^2_{SEN\_A_i}} * \sqrt{VV^2_{SEN\_B_i}}}\right)^2$$

Step 3: Integrate Sentence Similarity

$Sim_{Si\ Q} = q * (NC_{Si\ Q}) + (1-q) * (VC_{Si,Q})$

Where,

q = Balance coefficient

Algorithm 2: Identify_Indirect_Reviews

Input: The keyword 'k'

Output: A set of indirect review R' = {r1,r2,…, rn}.

Procedure:

1: Extract URL related to keyword
U = URL(k) = {U1,U2,…, Un} /* 'U' is an set of URL*/

2: For each URL Uiϵ U

3: Extract web data belongs to that URL and stored into file
D = Extract(Ui) /* 'D' is an stored web page data*/

4: For each sentence S ϵ D

5: Calculate Semantic Similarity between sentence and query
S' = Semantic_Similarity(S,Q)

6: If (S' >th) /* 'th' is threshold value */

7: Extract sentence and stored into text file
IR = {Sn1,Sn2,…, Snn} /* Set of Indirect Reviews*/

8: END If

9: End For

10: END For

---

5) Aspect Identification and Clustering:

To identify aspects POS tagger is use. The review which contains the noun or noun phrase is considered as a feature word. After getting aspects clusters of those aspects are formed using both bag of noun (BON)[15] and semantic similarity method. To identify semantic similarity Wu-Palmer algorithm is use.

Algorithm 3: Cluster_Reviews

Input: Set of direct and indirect reviews R' = {r_1,r_2,…, r_n} and predefine words dictionary D = {d_1,d_2,…, d_n}

Output: A set of clusters C = {c_1,c_2,…, c_n}.

Procedure:

1: For each sentence S ϵ R'

2: For each noun i.e n ϵ N
3: If (n.indexOf(S)) /*Check whether noun is part of sentence or not */
4: Claculate semantic similarity between noun and predefine words
If (Semantic_Similarity(n, di) > Th)
5: ci = di /* Assign predefine word to cluster */
6: ci = ci U S /*Add sentence into cluster*/
7: END if
8: End if
9: End For
10:END For

---

6) Polarity Identification and Classification:

To identify polarity bag of words method is use. The training dataset contains the 4818 negative words and 2041 positive words. After getting polarity words, reviews are classified into three classes i.e. positive, negative and neutral.

IV. RESULTS

The results of utility identification system are shown in following figures. To evaluate results mobile reviews dataset was used. The dataset contains 300 reviews on different aspects of mobile. Figure 3 shows the precision and recall values for five aspects of mobile. Precision and recall are calculated using following formulas:

$$Recall = \frac{A}{A+B} * 100$$

$$Precision = \frac{A}{A+C} * 100$$

Where,

A = Number of relevant reviews retrieved.
B = Number of relevant reviews not retrieved.
C = Number of irrelevant reviews retrieved.

Figure 4 shows purity values for clusters. The purity is an external evaluation criterion for cluster quality. It is the percent of the total number of objects that were classified correctly, in the unit range [0..1].

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j |c_i \cap t_j|$$

Where,

N = Number of reviews.
k = Number of clusters.
C_i = Cluster in 'C'.
t_j = Classification which has the max count for cluster 'c_i'.

Figure 5 shows the comparisons between final rating to the aspects of mobile by considering only direct reviews and both direct and indirect reviews.

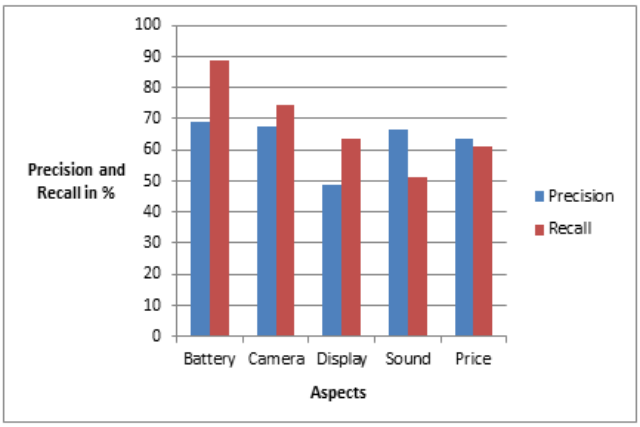

Fig. 3 Direct Reviews Extracted Using Cosine Similarity.

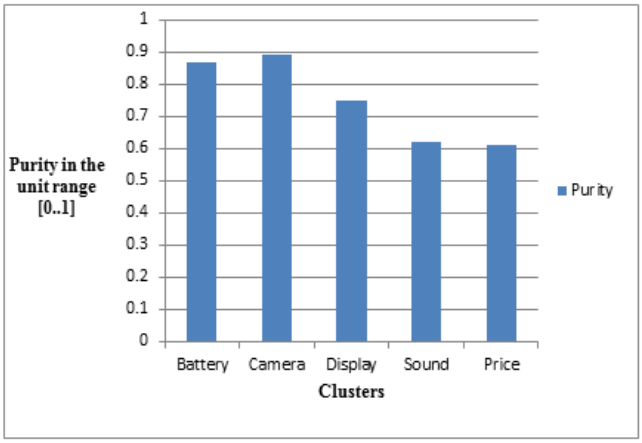

Fig. 4 Goodness of Clustering.

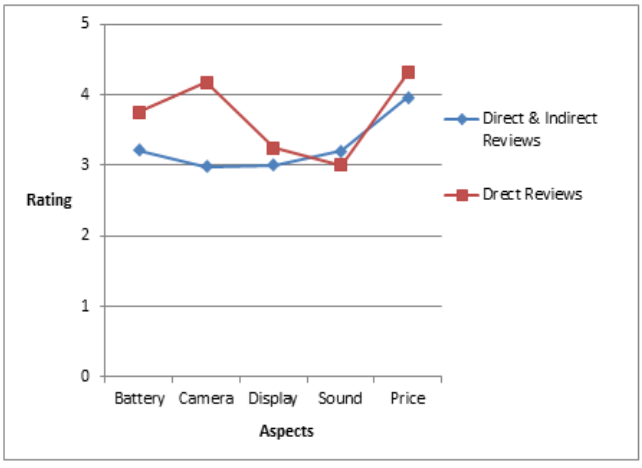

Fig. 5 Rating of Direct Reviews vs. Direct & Indirect Reviews.

V. CONCLUSIONS

In recent years, there is tremendous growth in social media, the web, blogs which help customers to post their opinion about the product on these websites. For an average customer, it is not possible to read all the reviews present on the different website, due to which there is chance that a customer may skip important reviews. So instead of considering reviews from the single website system should consider across the web. Hence, corpus based sentiment analysis is important to analyze sentiment. So the proposed system made use of web corpus for extracting indirect reviews based on a given keyword. The conceptual similarity of the term is used to extract meaningful reviews using well known, well studied Wikipedia and WordNet ontologies.

The experimental results show that average precision and recall values for direct reviews on 300 mobile review dataset are 62.45% and 67.93% respectively. The rating is calculated by considering direct and indirect reviews using ontologies shows the impact on different aspects.

REFERENCES

[1] Michael Gamon, Anthony Aue, Simon Corston Oliver, and Eric Ringger, "Pulse: Mining Customer Opinions from Free Text", in Advances in Intelligent Data Analysis. Springer 2008.

[2] Muflikhah L., Baharudin B.," Document Clustering Using Concept Space and Cosine Similarity Measurement", IEEE, Computer Technology and Development, pp. 58 – 62, 13-15 Nov. 2009.

[3] Wartena C., Brussee R., Slakhorst W., "Keyword Extraction Using Word Co-occurrence", IEEE, Database and Expert Systems Applications (DEXA), pp. 54 – 58, Sept 2010.

[4] Jingbo Zhu, Huizhen Wang, Muhua Zhu, Benjamin K. Tsou, Matthew Ma, "Aspect-Based Opinion Polling from Customer Reviews", IEEE Transactions on Affective Computing, vol. 2, no. 1, January-March 2011.

[5] XU Xueke, CHENG Xueqi, TAN Songbo, LIU Yue, SHEN Huawei, "Aspect Level Opinion Mining of Online Customer Reviews", IEEE, Communications, China (Volume:10 , Issue: 3), pp. 25 – 41, March 2013.

[6] Raisa Varghese, Jayasree M, "Aspect Based Sentiment Analysis using Support Vector Machine Classifier", IEEE, Advances in Computing, Communications and Informatics (ICACCI), pp. 1581 – 1586, 22-25 Aug. 2013.

[7] Manning, C.D., Schutze H., " Foundations of Statistical Natural Language Processing.", The MIT Press, Cambridge, Massachusetts, 1999.

[8] Meila, M., Heckerman D., "An experimental comparison of several clustering and initialization methods.", Technical report, Microsoft Research ,1998.

[9] Goodman J., "A bit of progess in language modelling", Technical report, Microsoft Report, 2000.

[10] Nyein S. S., "Mining contents in Web page using cosine similarity", IEEE, Computer Research and Development (ICCRD), pp. 472 – 475, 11-13 March 2011.

[11] Muflikhah L., Baharudin B.," Document Clustering Using Concept Space and Cosine Similarity Measurement", IEEE, Computer Technology and Development, pp. 58 – 62, 13-15 Nov. 2009.

[12] Wartena C., Brussee R., Slakhorst W., "Keyword Extraction Using Word Co-occurrence", IEEE, Database and Expert Systems Applications (DEXA), pp. 54 – 58, Sept 2010.

[13] Wen Zhang, Yoshida T., Xijin Tang, "TFIDF, LSI and multi-word in information retrieval and text categorization", IEEE International Conference on Systems, Man and Cybernetics, pp. 108 – 113, 12-15 Oct. 2008.

[14] Ming Che Lee, "A novel sentence similarity measure for semantic-based expert systems," *Expert Systems with Applications*, vol. 38, Issue 5, pp. 6392-6399, 2011.

[15] Farhadloo M., Rolland E., "Multi-Class Sentiment Analysis with Clustering and Score Representation," *IEEE 13th International Conference on Data Mining Workshops (ICDMW)*, pp. 904-912, 2013.