© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Extracting Opinion Relations from Online Reviews Based on WAM

D. Menaka*, E. Saravana Kumar

Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, TamilNadu, India

ABSTRACT

Data mining - process of pattern discovering in large data sets in that the emerging field is sentiment analysis. Opinion mining is the study of analyzing the human's opinions, sentiments, and emotion towards the entities such as products, services. The main application of sentiment analysis is collecting the online reviews about the product, social networks informal text. The process of the opinion targets and the opinion words extraction and determining the relations between these words. In previous, the nearest neighbor rules approach was used, the disadvantage of this method was not suitable for long span sentences. The Word Alignment Model is proposed to extract the opinion words and opinion targets from the obtained reviews and the graph based co-ranking algorithm is used to detect the opinion relation through opinion relation graph. While compared with previous used methods, this novel approach effectively decreases the error probability. The results shows the algorithm effectively outperforms when compare to existing methods.

Keywords: Data Mining, Opinion Mining, WAM

I. INTRODUCTION

Data mining - an emerging field of computing the data in the large data sets and exploring the pattern relations to the text. It has several challenges; the recent and series challenge is opinion mining. Sentiment analysis is the process of predicting the people opinions, sentiments, reviews and emotion towards the entities such as products, services. The main intention is to manufacture the opinions for the products from the enormous websites. The user gathered data from social media websites had a several content which has the impact about the topic [1]. The content provides the reviews extracted by the customer which is useful for the business persons.

Nowadays the large set of information available in the web, here the important issue is how to handle the data it's the major challenge. Hence the users have trouble to access those data [3]. The usage of internet in recent years are increased gradually, special process gives the data generated by user content. The main application of internet is e-commerce is affected by the user generated data. This needs the special mechanism to provide the better communicating data with uniqueness [5]. While considering the customers' point of view, when buying any product they need some reviews which provides the negative reviews or positive sentences. The major problem is the user should need to access the billions of opinions from the websites. Due to this the customer can make the decision to buy the products. Most of customers viewing for the lowest price to fulfill their requirements then they willing to buy the products.

With the explosive growth of social media for like blogs, snapdeal, flipkart. On the web, individuals and organizations are increasingly using the content in these media for decision making. Every site may have a huge amount of opinion about the product. The average human reader has the difficulty to find the most relevant reviews from online websites, then extracts and summarizes the opinion from different sites [7]. So automated analyzing the opinion systems is needed. The process of analyzing the opinion has classified by three levels.

The first level is document level, the task at this level is to classify whether an opinion of the document expresses a positive or negative sentiment. For example, the user can view a product review; the method detects the review whether it express a positive review or negative review about the product [9]. The second level is sentence level; the tasks at this level go to the sentences. In opinion search and retrieval and opinion question answering, sentences are usually retrieved and the sentences should be framed as opinion word and opinion target. Summarizing the opinion aims to select a set of sentences (or phrases) which summarize the opinion more accurately. While summarizing the opinion has to gather the set of reviews that summarizes the better opinion. The extracted review includes the several symbols, meaningless sentences; these should be removed by several methods [3]. The last level is aspect level, here both the document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level mainly determines the finer-grained process. Aspect level is known as feature level. Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level process at the opinion itself [9]. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion).

The main process of extracting the opinion word and opinion target from online reviews and providing the association relation among words can be done by Word Alignment Model (WAM) and Graph based Co-Ranking method [7]. In previous, the nearest neighbor rules and the syntax based method has the long span relation, error propagation. For optimizing process, the simulated annealing is used. Simulated annealing is based on meta-heuristic for global optimization problem to approximate the global optima in large space.

In this paper, section discusses about the related work and section 3 discusses detail about the methodology used for extraction and mining purpose and in section 4 specify about the SentiWordNet and section 5 discusses about the experiments and section 6 provides the conclusion of the paper.

II. RELATED WORK

In related work, a wide variety of techniques to support extraction of opinion words and opinion targets have been proposed in the literature. A general overview of each approach is presented prior to descriptions of specific techniques based on the existing approach. Literature Review is classified into three forms as, Document Level based Survey, Sentence Level based Survey, and Aspect Level based Survey.

Liu et al (2007) have proposed a sentiment based classification [12]. The main objective is identifying the opinion sentence from reviews and deciding whether each opinion sentence is positive or negative and summarizing the results. With the rapid expansion of e-commerce, more and more products are sold on the Web, and more and more 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 that they have purchased. With more and more common 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 grows rapidly. Some popular products can get hundreds of reviews at some large merchant sites. 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.

Zhu et al (2012) have proposed a Relational Adaptive bootstrapping (RAP) algorithm [9]. The objective is extracting the sentiment word from the text and generating the seed. Our basic idea is to first identify several common sentiment words across domains as sentiment seeds. Meanwhile, we mine some general patterns between sentiment and topic words from the source domain. Finally, we use the sentiment seeds and general patterns to generate topic seeds in the target domain. After generating the topic and sentiment seeds, we aim to expand them in the target domain to construct topic and sentiment lexicons. In this section, we propose a new bootstrapping-based method to address this problem. Bootstrapping is the process of improving the performance of a weak classifier by iteratively adding training data and retraining the classifier.

The Syntax based method to capturing the relation and ranking the product was proposed by zang et al (2010). Double propagation assumes that features are nouns/noun phrases and opinion words are adjectives [17]. It is shown that opinion words are usually associated with features in some ways. Thus, opinion words can be recognized by identified features, and features can be identified by known opinion words. The extracted opinion words and features are utilized to identify new opinion words and new features, which are used again to extract more opinion words and features. This propagation or bootstrapping process ends when no more opinion words or features can be found.

Chen et al (2011) have introduced a Word trigger method (WTM) to suggest tags according to the text description of a resource [10]. With the Internet, an increasing number of people are writing reviews. As a consequence, the number of reviews that a product receives grows rapidly. Some popular products can get hundreds of reviews at some large merchant sites. This makes it very hard for a potential customer to read them to help him or her to make a decision on whether to buy the product. In this research, we propose to study the problem of feature-based opinion summarization of customer reviews of products sold online.

Etzioni et al (2007) have proposed a Word Semantic Orientation [13]. The main objective is identifying product features and determines the polarity of opinions. While many other systems have used extracted opinion phrases in order to determine the polarity of sentences or documents, OPINE is the first to report its precision and recall on the tasks of opinion phrase extraction and opinion phrase polarity determination in the context of known product features and sentences. The datasets CRD and Large are used. OPINE extracts explicit features for the given product class from parsed review data.

Wang et al (2008) have proposed an Iterative Learning Method [11]. The task of identifying product features with opinion words and learning opinion words through features alternately and iteratively. In customer reviews, features and opinion words usually co-occur frequently, features are usually modified by the surrounding opinion words. If the absolute value of the relative distance in a sentence for a feature and an opinion word is less than Minimum-Offset, they are considered context dependent. The context-dependency property indicates the context association between product features and opinion words.

Zhang et al (2010) have described a Structure Aware Model Conditional Random Fields [17]. The process of summarizing the review based on document level extraction and extracts positive opinions, negative opinions and object features for review sentences. First, it can employ rich features for review mining. We will analyze the effect of features for review mining in this framework. Second, the framework can utilize the relationship among object features, positive opinions and negative opinions. It jointly extracts these three types of expressions in a unified way.

Hence the above paper discusses about the various methodology used for the extraction process and for mining the opinion relation.

II. METHODS AND MATERIAL

A. Word Alignment Model (WAM)

Word Alignment Model - monolingual method, this detects the opinion word and opinion target from online reviews for products. Opinion target is a noun/noun phrases which is used to express the users opinion about the product. Opinion word is a verb or adjective, where the opinion about the product has expressed.
For example [7],

"This phone has an amazing and colorful screen"

The opinion word and opinion target has extracted by WAM. In above example, the "screen" is an opinion word and the "colorful" and "amazing" is an opinion target [1].

The WAM method Liu et al (2015) proposed the following constrains:

- Nouns/noun phrases should be aligned with adjectives/verbs/a null word.
- Other unrelated words, such as prepositions conjunctions and adverbs should be aligned only with themselves.

The optimization process was done by simulated annealing. Simulated annealing is the process of optimizing the global optima based on the meta-heuristic method. In previous, the hill climbing algorithm was used. It has the main disadvantage; it suffers from getting stuck in local minima/maxima [19]. This can be overcome by simulated annealing; here it solves this problem by allowing worse moves to be taken some of the time. This method provides some uphill steps; hence this can be escape from the local maxima/minima.

Simulated annealing chooses random move from neighbor if the move is better than its current position then simulated annealing will ***always*** take it. If the worse move is found, then it will accept that move based on the probability. The below explained algorithm used for the optimization process

Algorithm:

Input: Review sentences Si = {w1; w2; . . . ; wn)
Step 1: Initialization: Calculate the word alignment
Step 2: Optimize toward the constraints
Step 3: Si= a;
Step 4: Loop do
    L= Neighbors(si);
    a1= NULL
Step 5: for all X in L
Step 6: if(a(X) > a1)
    a=X;
    a1= a(X);
Step 7: if a1 <= a(Si)
    Return Si;
    Si=x;
Step 8: end
Output: The calculated alignment for sentences.

B. **Graph Based Co-Ranking Algorithm**

The process of extracting the opinion word and opinion target obtained in previous step. Later, the relation between those words should be constructed by the opinion relation graph. Graph co-ranking method is estimated by candidate confidence of each opinion word and opinion target and this can be constructed on the graph [7]. The word which has higher problem will be extracted as opinion word or opinion target.

The candidate confidence can be estimated by random walking method. Here the confidence of an opinion target candidates and opinion word candidates in the iterations, then the higher confidence than the threshold are obtained as an opinion word or opinion target. The previous bootstrapping method has the error propagation problem. The graph based co-ranking algorithm effectively decreases the error problem.

Liu et al (2015) have provides the following features are used to represent the candidates:

- Salience feature: This feature indicates the salience degree of the candidates.
- Domain relevance feature: The opinion targets are domain specific and the difference between them has different domains.

In this process, we correct high-degree vertices to weak the collision and decrease the probability of a random walk running into unrelated regions on the graph [12]. Meanwhile, we calculate the prior knowledge of candidates for indicating some noises and incorporating them into our ranking algorithm to make combine operations on candidate confidence estimations. Finally, candidates with higher confidence than a threshold are extracted.

Instead, the confidence of each candidate is estimated by the global random walk process with the graph co-ranking. Intuitively, the error propagation is effectively alleviated [18]. The WAM should not force to define the alter relations to the displayed window; therefore it can capture more complex relations, such as long-span modified relations. While comparing with the syntactic pattern, the WAM is more robust because it does not need to parse informal texts. These opinions should find that equivalent modifier through word alignment.

III. SENTI WORD NET

The SentiWordNet is an extension for WordNet, such that all synsets can be associated with a value concerning the negative, positive or objective. This extension labels with some value to assign that in different category in between 0.0 to 1.0 [20].The sum of the three values is always 1.0, so each synset can have a nonzero value for each sentiment, since the synsets can have a positive, negative or objective depending on the context in which they are used. The web interface permit the user to look for the synset belongs to WordNet to associate with SentiWordNet values.

The system was built in two main steps. The first step is semi supervised learning and the second step is random walk. Then averaging of all given correlation results value between 0.0 and 1.0 can be obtained for each category for each synset. When classifiers determine the specific category, this sentiment will have the maximum value, which is 1.0. In the semi supervised step the training set for each of the out coming 8 classifiers is created. In the beginning small set of synsets is labeled physically and developed by automatic annotation. This can be quite efficient and helps to avoid labeling errors [22]. By having the training sets are classified with two learning methods. Seven positive and negative occurrences are selected from WordNet by human resources. These are very clearly neither negative nor positive. The set of positive and negative synsets are developing with the help of the WordNet relations. One of them can classify the synsets as either being positive or not positive; the other can classify them as neither being negative or not negative. If a synset is classified as either both positive and negative or not positive and not negative it is covered with the objective. The out coming classifier is a ternary classifier, able to classify the synset to be negative, objective or positive. It is registered to the whole WordNet to classify each synset.

The random walk algorithm is a graph based model that gives more weight to those nodes gaining the incoming links. The process is running iteratively on WordNet as a graph. The basic acceptance for this step is the following: if two synsets shares the ordinary context, then it's probably having the identical opinion. This is realized by "walking" from synset to synset and differentiating the relationship to its neighbours. If a synset occurs in the gloss of another synset they swing to have the identical polarity. So by this contrast, links of positivity and negativity are set up between the synsets [22]. When having more positive links, then they are tip to a certain synset, the higher its positive value will be in the end. So you can see positivity and negativity links between all synsets. Those that have a high number of positive or negative incoming links will have a greater value of positivity or negativity. As the values were very small later the random walk process, the final scores were normalized.

IV. RESULTS AND DISCUSSION

A. Datasets and Evaluation Metrics

For extracting the opinion word and opinion target, the main three datasets are chosen to analysis the WAM method. The dataset used here, CRD, COAE, and Large [7]. The CRD datasets contains the reviews about for five products. The COAE dataset contains the Chinese language reviews for the products like camera, car, laptop, and phone. The last dataset large contains the reviews on domains involves restaurant, hotel, mp3 [11], [15].

Three annotators are used in the annotation process for WAM and the two annotators were involved to extract whether noun/noun phrase is an opinion target or not. Here the conflict may occur. Then the third annotator is used to extract the final results. The previous methods like nearest-neighbour [5], syntactic pattern [3], double propagation, Word Translation Model [4] were also used in the same datasets. The three evaluation metrics are selected. The metrics are precision (P), recall (R) and F-measure (F) [1], [8].

Here the conflict may occur. Then the third annotator is used to extract the final results. The previous methods like nearest-neighbour [5], syntactic pattern [3], double propagation, Word Translation Model [4] were also used in the same datasets. The three evaluation metrics are selected. The metrics are precision (P), recall (R) and F-measure (F) [1], [8].

Precision is the percentage of selected items that are correct and the recall is the percentage of the correct items that are selected. A combined measure that assesses the precision and recall trade-off is the F-measure.

This model achieves high precision and recall. The graph based on co-ranking method constructs the opinion relation graph. Table 1 and table 2 specify the precision and recall for mining the opinion relation.

TABLE 1
PRECISION TABLE FOR OPINION RELATION

Method	**Laptop Precision**	**Phone precision**	**camera precision**
WAM	0.62	0.66	0.72
SP	0.53	0.57	0.64
WTM	0.45	0.50	0.55

TABLE 2
RECALL TABLE FOR OPINION RELATION

Method	**Laptop Recall**	**Phone Recall**	**camera Recall**
WAM	0.69	0.74	0.78
SP	0.58	0.62	0.67
WTM	0.50	0.58	0.61

V. CONCLUSION

Due to the high usage of internet, the extraction of huge volume of reviews about a product from the online websites to clarify the users taught is increasing day by day. To overcome this problem, the extraction of words and targets and providing relation among these words were followed. This process can be useful to the customer to clarify their needs. These processes has implemented by WAM and Graph based Co-Ranking algorithm and achieves the higher precision when compare to previous methods.

VI. REFERENCES

[1] Gan K. W. and Wong P. W. (2000), "Annotating information structures in chinese texts using hownet," in Proc. 2nd Workshop Chin. Lang. Process.: Held Conjunction 38th Annu. Meeting Assoc. Comput. Linguistics, Hong Kong, pp. 85–92.

[2] Gao Q., Bach N., and Vogel S. (2010), 'A semi-supervised word alignment algorithm with partial manual alignments,' in Proc. Joint Fifth Workshop Statist. Mach. Translation MetricsMATR, Uppsala, Sweden, pp. 1–10.

[3] Hai Z., Chang K., Kim J., and Yang C. C. (2014), 'Identifying features in opinion mining via intrinsic and extrinsic domain relevance', IEEE Trans. Knowledge Data Eng., vol. 26, no. 3, Pp. 623–634.

[4] Hu M. and Liu B. (2004), 'Mining and Summarizing Customer Reviews', in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Seattle, WA, USA, pp. 168–177.

[5] Hu M. and Liu B. (2004), 'Mining opinion features in customer reviews', in Proc. 19th Nat. Conf. Artif. Intell., San Jose, CA, USA, pp. 755–760.

[6] Li F., Han C., Huang M., Zhu X., Xia Y., Zhang S., and Yu H. (2010), 'Structure-aware review mining and summarization', in Proc. 23th Int. Conf. Comput. Linguistics, Beijing, China, pp. 653–661.

[7] Liu K., Xu L., and Zhao J. (2015), 'Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model', IEEE Transactions on knowledge and data engineering, vol. 27, no. 3.

[8] Li F., Pan S. J., Jin O., Yang Q., and Zhu X. (2012), 'Cross-domain co extraction of sentiment and topic lexicons', in Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Jeju, Korea, pp. 410–419.

[9] Liu K., Xu L., and Zhao J. (2012), 'Opinion target extraction using word based translation model', in Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn., Jeju, Korea, pp. 1346–1356.

[10] Liu Z., Chen X., and Sun M. (2011), 'A simple word trigger method for social tag suggestion', in Proc. Conf. Empirical Methods Natural Lang. Process., Edinburgh, U.K., pp. 1577–1588.

[11] Liu K., Xu H. L., Liu Y., and Zhao J. (2013), 'Opinion target extraction using partially-supervised word alignment model', in Proc. 23rd Int. Joint Conf. Artif. Intell., Beijing, China, pp. 2134–2140.

[12] Ma T. and Wan X. (2010), 'Opinion target extraction in Chinese news comments', in Proc. 23th Int. Conf. Comput. Linguistics, Beijing, China, pp. 782–790.

[13] Mei Q., Ling X., Wondra M., Su H., and Zhai C. (2007), 'Topic sentiment mixture: modeling facets and opinions in weblogs', in Proc. 16th Int. Conf. World Wide Web, pp. 171–180.

[14] Qiu G., Bing L., Bu J., and Chen C. (2011), 'Opinion word expansion and target extraction through double propagation', Comput. Linguistics, vol. 37, no. 1, pp. 9–27.

[15] Wang B. and Wang H. (2010), 'Bootstrapping both product features and opinion words from chinese customer reviews with crossinducing', in Proc. 3rd Int. Joint Conf. Natural Lang. Process., Hyderabad, India, pp. 289–295.

[16] Zhang L., Liu B., Lim S.H., and O'Brien-Strain E. (2010), 'Extracting and ranking product features in opinion documents', in Proc. 23th Int. Conf. Comput. Linguistics, Beijing, China, pp. 1462–1470.

[17] Zhao W. X., Jiang J., Yan H., and Li X. (2010), 'Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid', in Proc. Conf. Empirical Methods Natural Lang. Process., Cambridge, MA, USA, pp. 56–65.

[18] Zhu j., Wang h., Tsou B. k., and Zhu M. (2009), 'Multi-aspect opinion polling from textual reviews', in Proc. 18th ACM Conf. Inf. Knowl. Manage., Hong Kong, pp. 1799–1802.

[19] Purushothama G. K and Lawrence Jenkins," Simulated Annealing With Local Search—A Hybrid Algorithm for Unit Commitment,"IEEE Transactions On Power Systems, February 2003, Vol. 18, No. 1.

[20] Stefano, Andrea and Fabrizio"SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in Proc. 23th Int. Conf. Comput. Linguistics, Beijing, China, pp. 1462–1470.