International Journal of Advances in Computer and Electronics Engineering
Volume: 1 Issue: 1, May 2016, pp. 35 – 39

Optimized Summary Generation Using Genetic Algorithm

[1]P. Showmiya, [2]V. Priya
[1]PG Scholar, Department of Computer Science and Engineering,
Dr. Mahalingam College of Engineering and Technology, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Science and Engineering,
Dr. Mahalingam College of Engineering and Technology, Tamil Nadu, India
[1]pshowmiya65@gmail.com, [2]priya@drmcet.ac.in

Abstract: *In today's world, the volume of reviews available in web is increasing exponentially. So, it is more challenging to analyze such huge amount of reviews. These reviews are available in the form of reviews, comments and feedbacks. To utilize this information effectively, Optimization Technique is used. The important aspects are identified by the number of occurrences of keywords in various customer reviews. The obtained aspects along with its opinion are passed to Naïve Bayes Classifier to classify the aspects under the sentiment terms such as positive, negative or neutral. These aspects are ranked using Probabilistic Ranking Algorithm which is used to infer the importance of various aspects. So the approach of optimized summary generation using genetic algorithm is expected to improve the overall summary.*

Keyword: *Aspect ranking; Sentiment classification; Aspect Identification; Optimization summarization;*

1. INTRODUCTION

Opinion mining is a Natural Language Processing (NLP) and Information Extraction (IE) task that aims to obtain the feelings of an author expressed as positive or negative opinions by analyzing a large number of documents. Opinion mining involves the methods from computational linguistics and Information Retrieval techniques.

Opinions could be expressed on anything such as reviews, blogs, discussion groups, forums etc. Large numbers of such opinions prevail in online platforms. Opinion mining helps to analyze this content and gives the summary of the overall opinions. The usage of Sentiment analysis techniques are abundantly growing in the commercial environment.

In common, the overall contextual polarity or sentiment of an author, near a particular aspect can be determined using sentiment analysis. The key challenge in this area is the sentiment classification in which the sentiment might be a sentence, or evaluation of a thing namely film, book, product, etc. This can be in the form of a document or a sentence or a feature that can be considered as positive or negative .Classifying entire documents according to the opinions towards certain objects is called as sentiment classification.

Analysis of sentiments may be document based where sentiment in the entire document is summarized as positive, negative and neutral comments. Customer reviews can be sentiment oriented, appraisal oriented or emotions of the particular topic towards entities such as products or organizations. Internet aids exchange of public opinions with respect to some topic. It is very important factor that sentiment analysis analyzes the data from the public opinions but identifying those sentiments are difficult. After sentiment analysis process, ranking is to be performed. Product aspect ranking is valuable to wide variety of applications. It performs extensive evaluations on products and demonstrates the potential of aspect ranking. Extractive review summarization in text is summarized for aspect. Probabilistic aspect ranking algorithm is used which effectively exploits the aspect frequency as well as the influence of opinions given to each aspect. The review is generated based upon the weighted aggregation of the opinions.

However, the proposed framework determines the summary by using Genetic algorithm for providing an optimized summary. Genetic Algorithm is an Optimization Technique and the aspects under genetic algorithm are assigned a fitness based on the Multi Objective Function. It has a tendency to find solutions near the individual best solution of its objective.

Genetic Algorithms (GAs) are adaptive heuristic search algorithm created on the evolutionary thoughts of natural selection and genetics. As such they characterize an intelligent exploitation of a random search used to solve optimization problems. Even though randomized, GAs are by no means random, instead they exploit historical facts to direct the search into the region of improved performance within the search space. Opinion summarization is the process which gives the summary of overall opinions on the aspect in the review. It does not give summary by simply selecting some subset of the data or rewriting the sentences as such.

The rest of this paper is organized as follows: Section 2 describes the related work. Later in Section 3, Existing system has been explained briefly. The proposed system has been briefly explained in Section 4. Section 5 has a discussion done on the dataset used and experiments conducted. Finally, Section 6 discusses the conclusion of this paper.

2. LITERATURE SURVEY

The Aspect based opinion mining methods divide input texts into aspects, also called features. Hu's work [4] can be considered as a pioneer work on feature-based opinion summarization where frequently occurring noun and noun phrases are considered as aspects. Nathan's work [12] describes the approach ISODATA clustering. It also uses a number of different heuristics to determine whether to merge or split clusters. Points are assigned to their closest cluster centers and cluster centers are updated to be the centroid of their associated points. Clusters with very few points are deleted, large clusters satisfying some conditions are split, and small clusters satisfying other conditions are merged.

Manevitz [13] determines the approach to one class SVM's for document classification. The SVM approach as represented by Scholkopf was superior to all the methods except the neural network. However, the SVM methods turned out to be rather sensitive to holder more data but they are useful to handle positive information. Yogesh Kumar's work[14] employed the methodologies using Genetic Algorithm. This paper gives a review of the growth in the techniques of text Summarization. They obtain broad set of features in which they apply the features in the fitness function.

Elena Lloret [16] presented an idea about Text Summarization in Natural Language Processing. This speaks about Automatic text summarization, Extraction and Abstraction. Different types of inputs under various domains like Biography, news etc. have been subjected to feature extraction and its corresponding outputs are also been described shortly.

Text summarization is classified into two broad categories namely abstraction and extraction. Summary extraction was proposed by Archana AB, Sunitha C [17] concentrated on Query-focused summary extraction to find more relevant documents accurately based on the query fed by the users. Four different techniques namely Neural Network, Graph Theoretic, Fuzzy based method and Cluster based method have been studied and compared.

A real time problem has been addressed by Sunita R. Patil and Sunita M. Mahajan [18] which addresses summarization using data mining techniques like Sentence Clustering, Scoring, Summary optimization etc. Data mining strategies such as extraction and clustering are used for finding 'Research Relevant Novel' (RRN) terms. Mining relevant sentences from multiple text documents uses 'Maximal Marginal Relevance' (MMR) criteria holding RRN terms. A Query based opinion summarization which used LSI based method was proposed by Feng Jin, Minlie Huang, Xiaoyan Zhu [19]. The query that the user provide must be an opinion based query. When the query is provided the system finds it very easy to identify user's intention and return relevant summary.

Lun-Wei Ku, Yu-Ting Liang et al. [20] concentrated on opinion extraction, summarization and proposed algorithms which extract text in word, sentence and document level. Involving topical words enhances the performance of opinion extraction. Hyun duk kim, kavita ganesan et el. [21] compared various classification techniques like Sentiment Classification, Subjectivity Classification, Text Summarization and Topic Modeling. All pros and cons of the above mentioned classification techniques are clearly addressed. Customer reviews are given as the input and preprocessed in which POS Tagging is done and aspects are identified from the reviews. Naïve Bayes Classifier is used for sentiment classification and finally they are ranked using Probabilistic Ranking Algorithm. Based on the score Extractive summarization technique is used to mine the most informative segments (e.g. sentences or passages) from the source reviews.

2.1. Inferences from Existing System

Extractive review summarization does not improve the accuracy performance of summarization, because it is only formulated with the help of informative sentences whereas Optimization Technique is expected to improve the aspects of overall summary

3. PROPOSED SYSTEM

System architecture describes the overall flow of summarization. The frequent noun terms are extracted with the help of Stanford parser and later Naïve Bayes classifier is used to classify the aspects whether it is good, bad or neutral. After sentiment classification Ranking algorithm is used to infer the important aspects and then summarization is to be generated. Optimization technique is formulated where overall summary that is generated is expected to improve the accuracy performance.

3.1 Aspect Identification

Aspects do not directly appear in a text but they appear in the manner of aspect expressions. Initially, the dataset in the form of reviews are collected for each products and the dataset is preprocessed by removing the stop words from each of the sentences to reduce the noise. Once the stop words are removed, Parts-of-speech tagging is done for each word. Parts-of-speech tagger uses the Stanford parser which parses each sentence and yields the part-of-speech tag of each word (whether the word is a noun, adverb, verb etc.) and identifies simple noun and verb groups. Each sentence in the review database is stored along with the POS tag information of each word in the sentence.

POS Tagger identifies the aspects by extracting the frequent noun terms in the reviews. First, it identifies the nouns and noun phrases in the documents. Second, the occurrences of the nouns and noun phrases are counted, and only the frequent ones are kept as aspects.

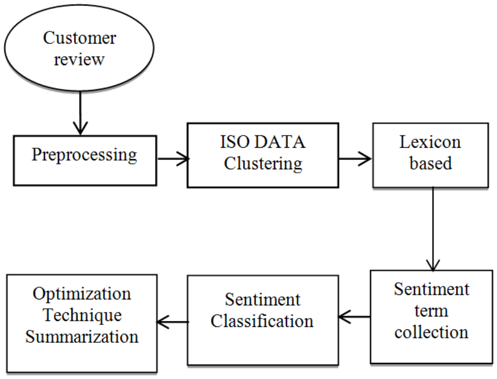

Figure 1 System Architecture

3.2 Sentiment Classification

The lexicon-based systems use a sentiment lexicon containing a list of sentiment words, phrases and idioms to determine the sentiment orientation on each aspect. On the other hand, the supervised learning methods train a sentiment classifier constructed on training corpus. The classifier is then used to predict the sentiment on each aspect. In this work, the customer reviews have been explicitly categorized into positive and negative opinions Based on the aspects. Specifically, it collects the sentiment terms from customer reviews based on the sentiment lexicon provided by the classifier. Here, the technique called Naïve Bayes Classifier is used to determine the opinion on the opinionated expression.

3.3 Aspect Ranking

Once the sentiment is classified, the next process is to rank the aspects using Probabilistic Aspect Ranking Algorithm. At first, the algorithm produces the scores for each aspect and then it identifies the important aspects of a product based on that generated scores. Finally, the overall rating in each review is formulated based on the weighted sum of the opinions on particular aspects.

3.4 Optimization Summarization

Genetic Algorithm is parallel to the process of natural evolution in order to optimize linearly search problems. The operators used in Gas are selection, crossover and mutation. The problem space is represented in the form of chromosomes. Optimization is used to improve the accuracy of Summarization technique. It is the simplest possible multi-objective GA derived from Multi-Objective Optimization. Here fitness evaluation function is used for five objective functions. Each individual in the first subpopulation is assigned a fitness based on the second objective function. This is particularly useful in handling problems where objective function take values of different orders of magnitude.

In objective function, subpopulation determines the good solutions corresponding to the particular objective function. No two solutions are compared for different objective functions. It does not create any proportionate selection operator. It performs proportionate selection operator to create a mating pool. GAs iteratively updates a population of individuals. Individuals in the population are represented by bit strings. The fitness function evaluates with the help of representation of chromosomes. The GA must be prepared with a fitness function allowing it to score and to rank the individuals.

Algorithm:

```
Optimization (n, x, μ)
   // Initialise generation 0:
   k := 0;
   Pk := a population of n randomly-generated individuals;
   // Evaluate Pk:
   Compute fitness (i) for each i ∈Pk;
   do
   { // Create generation k + 1:
   // 1. Copy:
   Select (1 − x) × n members of Pk and insert into Pk+1;
   // 2. Crossover:
   Select x × n members of Pk; pair them up; produce offspring; insert the offspring into Pk+1;
   // 3. Mutate:
Select μ × n members of Pk+1; invert a randomly-selected bit in each;
   // Evaluate Pk+1:
   Compute fitness i) for each i ∈Pk;
   // Increment:
   k := k + 1;
   }
   While fitness of fittest individual in Pk is not high enough;
return the fittest individual from Pk;
```

Here crossover and mutation is applied to create a new offspring. In chromosome representation, it is easy for us to encode one of our table-driven agents as a bit string and just as easy to decode a bit string to recreate a table-driven agent. The fitness function used in Optimization summarization is

$$\text{Score}(S_i) = \sum_{j=1}^{NooofFeatures} \left(x_j \times f_i(S_i)\right)$$

The fitness function is composed of various features, for the features crossover and mutation is applied to get the values of weights. After an initial population is randomly generated, the algorithm evolves through three operators:

- Selection which equates the fittest solution.
- Crossover represents the mating between individuals

- Mutation is performed with random values.

The fitness function is generated for the particular ith sentence for which score is being calculated. n is the number of individuals in the population; x is the fraction of the population to be replaced by crossover in each iteration; and µ is the mutation rate.

The Population at any generation is divided into five equal divisions. Each individual in the first population is assigned as fitness based on first Objective function. In second population, the individuals are assigned as fitness based on second Objective function. The Optimization technique is particularly used in handling problems with the subpopulation is to be assigned.

4. EXPERIMENTAL RESULTS

4.1 Dataset

The dataset used for the experiments are the Product reviews. The product used is Canon G3. The existing system used the reviews which were which were given about a particular camera review. The summary is generated for each of the aspects according to their score Informative Sentence. The proposed system uses the same dataset and overall summary is to be generated. The overall summary is expected to improve the accuracy performance in it.

4.2 Discussion

The existing system ranks the aspects based on both the frequency and score importance of aspects in the review. Table 1 contains the different aspects about camera along with its score and rank.

TABLE I ASPECT RANKING

Aspects	**Sentiment Classification Score**	**Rank**
Clarity	97.764	1
Memory	83.415	2
Battery	80.254	3

The two measures that have been considered for summarization are Normalized discounted cumulative gain (NDCG) [1] and Recall Oriented Understudy for Gisting Evaluation (ROUGE) [1]. NDCG is a measure used widely to evaluate the performance of ranking. The formula to calculate the NDCG is given below.

$$\text{NDCG}@k=\frac{1}{Z}\sum_{i=1}^{k}\frac{2^{t(i)-1}}{\log\,(1+i)}$$

Where t(i) is the importance degree of the aspect at position i, and Z is a normalization term derived from the top-k aspects. Another widely used performance metric is ROUGE which is applied to evaluate the quality of the summary. The formula to calculate the ROUGE is given below.

$$\text{ROUGE}=\frac{\sum_{s\epsilon\{Reference\,summaries\}}\sum_{gram_{n}\epsilon s} Count_{match}(gram_{n})}{\sum_{s\epsilon\{Reference\,summaries\}}\sum_{gram_{n}\epsilon s} count(gram_{n})}$$

Where n stands for the length of the n-gram and count$_{match}^{(gram\ n)}$is the maximum number of n-grams occurring in the candidate summary. The summary that is generated using Optimization technique is expected to improve the accuracy performance. Overall summary is generated with the help of genetic algorithm.

5. CONCLUSION

The experimental results shows that the aspect ranking algorithm to identify the important aspects of products from frequent reviews. The algorithm that determines the aspect frequency and the influence of customer reviews given to each aspects over the opinions. The experimental corpus contains the 286 MB of data. Optimized summary is expected to improve the performance of summary by using genetic algorithm when compared to probabilistic ranking approach.

REFERENCES

[1] Zheng-JunZha, Jianxingyu, JinhuiTang,Tat-Seng Chua,(2014), ‘Product Aspect Ranking and its Applications’, IEEE Transactions on knowledge and data engineering, vol.26, No.5.

[2] ArtiBuche, Dr.M.B.Chandak, AkshayZadgaonk (2013), ‘Opinion Mining and Analysis: A Survey’, in Proceedings of the International Journal on Natural Language Computing, Volume 2, No. 3, pp. 39-48.

[3] Vinodhini.G and Chandrasekaran.RM.(2012) , ‘Sentiment Analysis and Opinion Mining: A survey’, in Proceedings of the International Journal of Advanced Research in Computer Science and Software Engineering ,Volume 2,No 6,pp. 282-292.

[4] Minquing Hu, Bing Liu (2005), ‘Mining and summarizing customer reviews’, in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and Data mining, pp.168 177.

[5] PopescuAnaMaria, OrenEtzioni (2005), ‘Extracting product features and opinions from reviews’, in Proceedings of the conference on Human Language Technology and Empirical methods in Natural Language Processing,pp. 339-346.

[6] KunpengZhang, Ramanathan Narayanan, Alok Choudhary (2010), ‘Voice of the Customers: Mining Online Customer Reviews for Product feature-based Ranking’, in Proceedings of the 3rd Conference on Online social networks. pp.1-11

[7] YongyongZhail, YanxiangChenl, Xuegang Hu (2010), ‘Extracting Opinion Features in Sentiment Patterns’, in Proceedings of the International Conference on Information Networking and Automation, Volume 1, pp. 115-119.

[8] Minquing Hu and Bing Liu (2004) , ‘Mining Opinion Features in Customer Reviews’, in Proceedings of the 19th national conference on Artificial Intelligence,Chicago,pp.760.

[9] Bing Liu, Mingquing Hu andJunsheng Cheng, (2005) ‘opinion observer: Analyzing and Comparing Opinions on the web’, WWW, ACM. Chiba Japan, pp.91-134.

[10] Bing Liu and Mingquing Hu, 7 (2006) ‘Opinion Extraction and Summarization on the Web’, Association for the Advancement of Artificial Intelligence (AAAI), pp.1621-1624.

[11] WeishuHu ,Zhiguo Gong, Jingzhiguo,(2010) ‘Mining Products Features from Online Reviews’, ‘Faculty of Science and Technology University of Macau, China,pp.24-29.

[12] Nathan S.Netanyahu,(2007), 'A Fast Implementation Of the Isodata Clustering Algorithm', International Journal of Computational Geometry and Applications. pp.71-103.

[13] Larry M.Manevitz, Malikyousef, (2011), 'One-class SVMs for Document Classification', Journal of Machine Learning Research pp.139-154.

[14] Vogeh Kummar Meena, Dinesh Gopalani,(2015), 'Evolutionary Algorithms for Extractive Automatic Text Summmarization', International Conference on Intelligent computing, Communication and Convergence, pp.244-249.

[15] Atif khan, Naomiesalim, (2014),'A review on abstractive summarization methods', Journal of Theoretical and Applied Information Technology Vol. 59 No.1.

[16] Elena Lloret, (2006), 'Text Summarization: An Overview', International Journal on Advanced Computer Theory and Engineering.

[17] Archana AB, Sunitha. C, (2103), 'An overview on Document Summarization Techniques', International Journal on Advanced Computer Theory and Engineering, 2319 – 2526, Volume-1, Issue-2

[18] Sunitha M. Mahajan, (2012), 'Optimized Summarization of Research Papers as an Aid for Research Scholars using Data Mining Techniques', International Conference onRadar, Communication and Computing, pp.369-375.

[19] Feng Jin, Minlie Huang, Xiaoyan Zhu,(2008),'A Query-specific Opinion Summarization System', Dept. Computer Science and Technology, Tsing University, China.

[20] Lun-Wei Ku, Yu-Ting Liang et al., (2006), 'Opinion Extraction, Summarization and Tracking', AAAI spring symposium: Computational approaches to analyzing weblogs. Vol. 100107.

[21] Kim, Hyun Duk, et al., (2011), 'Comprehensive review of opinion summarization'.