Particle Swarm Optimization Based Feature Selection and Summarization of Customer Reviews

B.Suganya,
PG Scholar, Department of Computer Science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi- 642002, India,
V.Priya, *M.E.,*
Assistant professor, Department of Computer Science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi- 642002, India

Abstract- The steady growth of e-commerce has led to a significantly large number of reviews for a product or service. This gives useful information to the users to take an informed decision on whether to acquire a service and/or product or not. Opinion mining techniques are used to automatically process customer reviews for extracting feature and opinion in a concise summary form. Existing feature based summarization system uses dependency relations and ontological knowledge with probabilistic based model to generate the summary. To enhance the accuracy of summarization, fitness proportionate binary particle swarm optimization (FBPSO) based feature selection is proposed. The BPSO could efficiently search for subset of features using fitness sum based on the multi-objective function. The FBPSO overcome the problems of traditional BPSO as it focus much on the overall performance of a particle as a whole and it does not pay attention on every single feature. The multi-objective function used in FBPSO is based on dominance, mutation and crowding factor to generate an efficient summary. The performance of the system is measured using the Recall-Oriented Understanding for Gisting Evaluation (ROUGE) toolkit. Experimental results show that the proposed approach of summary generation using multi-objective BPSO algorithm outperforms the traditional probabilistic model.

Keywords— *Feature selection, Multi-objective, Binary Particle Swarm Optimization, Summarization*

I. Introduction

The steady increase in number of e-commerce portals has made the lives of people easier in terms of the efforts spent for buying a product or acquiring any service. As more people seek the advices of fellow users for making any informed decision. Different customers or users may express their views on different aspects of the products, or services and, hence, the amount of information available is significantly large in size. In order to extract the most useful information, any customer or manufacturer has to read all the reviews that have been written. Unfortunately, this task is not simple because of the following facts: searching through the entire collection of reviews for any particular aspect is a hugely involved task that consumes lots of time and efforts, communication in these mediums is informal, and the texts are not very well-formed. Therefore, there is an urgent need to develop applications that aid in mining the desired information from this huge collection of online contents.

Opinion mining is a Natural Language Processing and Information Extraction task that aims to obtain the feelings of the writer expressed as positive or negative opinions by analyzing a large number of documents. The overall contextual polarity or sentiment of an author about some aspect can be determined using sentiment analysis. The main challenge in this area is the sentiment classification and summarization. Opinion mining helps to solve this problem and gives the summary of the overall opinions.

Opinion Summarization aims at creating concise summary from large number of reviews. It does not give summary by simply selecting some subset of data or rewriting the sentences as such. The simplest form of opinion summary is the result of sentiment prediction by aggregating the sentiment scores. Features based summarization extracts the features and then identify opinions that associate with it to generate a text based summary.

The main goal of this paper is to find a feature subset for a sentiment classification and summarization problem with PSO based methods. In order to make the selected feature subset as powerful as possible, there is a need to improve PSO’s performance in discrete space. In order to achieve this goal a new way to calculate velocities based on fitness values is then proposed and based on that new binary version of PSO called Fitness Proportionate Binary Particle Swarm Optimization.

This paper focus on a Particle Swarm Optimization based feature selection and summarization of customer reviews. It transfers all the texts into corresponding feature vectors, and then fitness proportionate based feature selection is performed. It is a concept often seen in genetic algorithms. It uses the fitness sum based on multi-objective function in the selection step. Multi-objective function involves minimizing or maximizing multiple conflicting objective functions. Then the features extracted using multi-objective optimization is used to generate the summary. The main advantage of using multi-objective optimization is to keep the diversity of the swarm and to improve the search ability of the swarm.

II. RELATED WORK

Bing Xue et al. [1] proposes two feature selection approaches, single feature ranking and Binary Particle Swarm Optimization (BPSO). In the first approach, individual features are ranked based on classification accuracy using only a few top ranked features. BPSO is applied to feature subset ranking to search different feature subsets. BPSO could efficiently search for subsets of complementary features to avoid redundancy and noise. When using BPSO to solve the feature selection problems, the representation of a particle is an n-bit binary string. Where n is the number of features and the dimensionality of the search space. The feature mask value of "1" represents the feature is selected and "0" otherwise. Experimental results show that BPSO based approach outperforms single feature ranking approach.

Sentiment classification makes feature selection as important for the purpose of summarization. Frequently used bag-of-words model (BoW) proposed by Wang et al. [8] produces a very large feature space due to diverse choice of words. This problem is solved by adopting an efficient feature selection process to remove the redundant features. In BoW model, the number of dimensions of each feature is the number of different words. Each feature represents one word and some words contain semantic orientations. If the semantic orientations of the words are considered, the feature selection result will be more efficient and are useful in subsequent operations.

Zhou Z et al. [9] proposed fitness proportionate based feature subset selection to make the selected feature subset as more powerful. To achieve this, two major problems of PSO including search space is discrete and traditional way of calculating velocity is considered. Similar to the genetic algorithm, every feature selection result is transformed into a bit string. The value of each bit represents whether the corresponding feature is selected or not. The fitness value of each particle is calculated and based on that it updates the local and previous best values. Based on these values the best feature subset is returned.

Mohammed Salem Binwahlan et al. [6] investigate the effect of the feature structure on the features selection using particle swarm optimization. The particle swarm optimization is trained to learn the weight of each feature. The features used are different in terms of the structure, where some features were formed as combination of more than one feature while others as simple or individual feature. Therefore the effectiveness of each type of features could lead to mechanism to differentiate between the features having high importance and those having low importance. The combined features have higher priority of getting selection more than the simple features. In each of the iteration, the particle swarm optimization selects some features, then corresponding weights of those features are used to score the sentences and the top ranking sentences are selected as summary. The selected features of each best summary are used in calculation of the final features weights.

PSO for feature selection process has a potential limitation of losing the diversity of the swarm quickly during the evolutionary process. In order to better address feature selection problem, Bing Xue et al. [2] proposed another PSO based multi-objective feature selection algorithm. The multi-objective algorithm is based on the ideas of crowding, mutation and dominance parameters. Crowding factor is employed to decide which nondominated solutions are to be added to leader set and kept during the evolutionary process. Mutation is adopted to keep the diversity of the swarm and to improve the search ability of the algorithm. A dominance factor is used to determine the size of the archive, which is the number of nondominated solutions the algorithm reports. Experimental results show that these three factors take advantage of the different behaviors to improve the search ability for the nondominated solutions.

Opinion extraction mines opinions at word, sentence and document levels from articles. Opinion summarization summarizes opinions of articles using sentiment polarities, degree and the correlated events. Lun-Wei Ku et al. [4] proposed an algorithm for opinion extraction at word, sentence and document level. The issue of relevant sentence selection is discussed, and then topical and opinionated information are summarized.

Dim En Nyaung et al. [3] proposed a task of opinion summarization. Product feature and opinion extraction is critical to opinion summarization, because its effectiveness significantly affects the identification of semantic relationships. The polarity and numeric score for all the features are determined using Senti-WordNet Lexicon. The problem of opinion summarization refers how to relate opinion words with respect to a certain feature regardless the distance from features to opinions.

III. PROPOSED SYSTEM

This section presents the architecture of the proposed system to identify the feature-opinion pairs for summarization. In Particle Swarm Optimization based feature selection and summarization of customer reviews, the dataset document reviews are represented as a feature vector during data preprocessing. After all the texts are transformed into corresponding feature vectors, Fitness proportionate selection is performed. It is a concept which is often seen in genetic algorithms. It selects the solution based on its fitness score on a particular task. The higher fitness score means the solution has the more chances have of getting selected because the solution is considered to be more suitable for the task. After extracting the features, opinion scores are assigned to them for the purpose of summarization.

Fig. 1 presents the complete architecture of the proposed system, which consists of the following stages: Dataset collection and preprocessing, feature selection, sentiment classification and summarization.

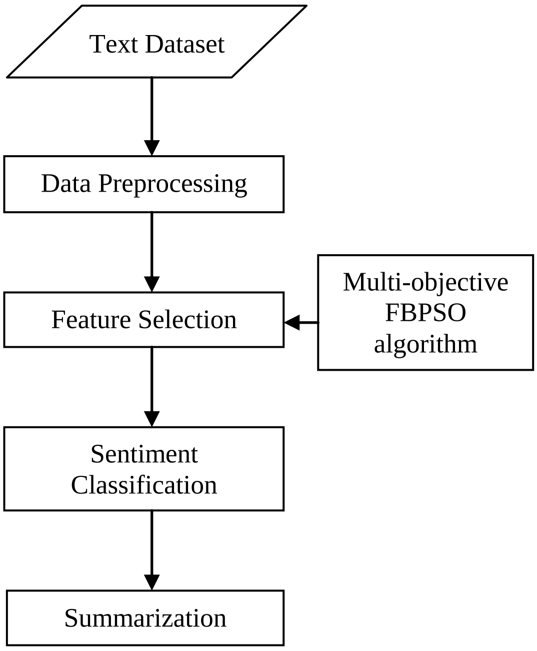

Fig 1: Proposed System Architecture

A. *Data Preprocessing*

Every document in the dataset is preprocessed in which word segmentation and stop words removal are done to reduce the noise. Document in the dataset is transformed under a specific model and represented as a feature vector during data preprocessing. One commonly used model for document representation is unigram bag-of-words model (BoW). Under BoW model, number of dimensions of each feature vector is the number of different words in the whole text dataset. The vector assigns "1" to dth dimension if the text contains corresponding word, and assigns "0" if it does not., After all the texts are transformed into corresponding feature vectors, sentiment classification is done.

B. *Feature Selection*

In classification tasks, the process of feature selection must be executed, in which a smaller set of all features is gained in order to achieve better classification performance. PSO can be used to perform feature selection only in continuous space. In order to use PSO to solve discrete problems, binary particle swarm optimization (BPSO) is proposed. Two changes are made to the continuous PSO. First is the representation of the particle's position. In continuous PSO where a particle's position is a set of real values, in BPSO, particle's position becomes a binary vector. The second change is that the velocity of a particle no longer represents the direction of a particle's movement. It means the possibility of choosing "1" at a specific bit of a particle's movement.

The velocity of a particle is updated at each iteration based on the fitness value in the fitness proportionate selection step. Fitness proportionate selection is a concept often seen in genetic algorithms. It selects the solution based on its fitness score. The higher the fitness score a solution has, the more chances the solution gets selected. The velocity and position of a particle is represented as follows:

$$v_{id}^{t+1} = \begin{Bmatrix} mr, & \text{if } n_0 == 0 \\ 1 - mr, & \text{if} n_1 == 0 \\ \frac{f_1}{f_1 + f_0}, & \text{otherwise} \end{Bmatrix} \quad (1)$$

$$x_{id}^{t+1} = \begin{Bmatrix} 0, & rand < v_{id}^{t+1} \\ 1, & \text{otherwise} \end{Bmatrix} \quad (2)$$

Where n_0 is the number of involved particles with $x_{id} = 0$ and n_1 is the number of involved particles with $x_{id} = 1$. f_1 and f_0 are computed as: First, the involved particles are divided into two sets S_1 and S_0 based on whether they select 0 or 1.Then, f_1 and f_0 are calculated by averaging the fitness values of particles in S_1 and S_0 respectively. The fitness value of the particle is given by,

$$f(x) = \sum_1^n p_i x_i - s * \max p_i \quad (3)$$

where

$$s = |j| \sum_1^m w_{ij} x_i > c_j \quad (4)$$

p_i is the value of particle i, w_{ij} is the weight of the particle i for swarm j, c_j is the size of the swarm, n is the number of particles and m is the number of constraints.

The detailed algorithm is shown in Fig. 2. For each particle, we select the leader by means of a binary tournament based on the crowding value of the leaders. The maximum size of the set of leaders is fixed equal to the size of the swarm (or population). After each generation, the set of leaders is updated, and so are the corresponding crowding values. If the size of the set of leaders is greater than the maximum allowable size, only the best leaders are retained based on their crowding value. The rest of the leaders are eliminated. The swarm is subdivided in two parts Each sub-part of the swarm will adopt a different mutation scheme: the first sub-part will have uniform mutation, the second sub-part will have non-uniform mutation. With the use of these different operators we are aiming to have the ability of exploring (uniform mutation) and exploiting (non-uniform mutation) the search space as the process progresses. The available set of leaders is the same for each of these sub-parts. Additionally, each particle can use as a leader a particle produced by a different sub-part of the swarm. In this way, the different sub-parts of the swarm will share their particular success and the final results will be a combination of using different behaviors inside the same swarm.

```
FBPSO based Feature Selection
Initialize parameters of BPSO
Randomly initialize swarm
initialize the set of leaders LeaderSet and Archive
calculate the crowding distance of each member in LeaderSet
while stopping criterion not met do
    calculate each particle's fitness function
    for i=1 to swarmSize do
        update the localbest of $p_i$
        update the previousbest of $p_i$
    end
    for i=1 to swarmSize do
        for j=1 to dimension do
            update the velocity of $p_i$ according to equation (1)
            update the position of $p_i$ according to equation (2)
        end
    end
end
Return the best feature subset found by the swarm
```

Fig 2: FBPSO based Feature Selection

C. Sentiment Classification

Sentiment classification has some special traits and characteristics different from traditional classification problems. Each feature in sentiment classification model refers to a unique word. Most words in a language contain their semantic orientations from the linguistic perspective so it is reasonable to conjecture that the semantic orientation of the word feature itself will help the selection of a proper feature subset. A sentiment lexicon consists of a large amount of expressions and their sentiment polarities. A sentiment lexicon is utilized as the resource of linguistic knowledge in the sentiment classification-oriented feature selection. If a word is found in the lexicon, it means the word itself contains sentiment polarity and is more likely to be a useful feature in sentiment classification. According to previous studies, the feature matrix in sentiment classification is often extremely large and sparse due to authors' diverse choice of words. It is highly possible that many of those word features will provide similar functions to each other. So the degree of redundancy should also be considered in sentiment classification-oriented feature selection schemes. Based on mutual information, it is possible to evaluate the contribution of a single feature in the feature subset. If f is the target feature in feature subset S, and the contribution that f makes to S is calculated by subtracting the average mutual information between f and all members in S from the mutual information between f and class label C. The contribution measure not only focuses on the usefulness of feature f itself, but also pays much attention to the redundancy between selected features by subtracting their shared information. By using the contribution measure, redundant features are more likely to be removed from the selected feature subset, which can make the final subset smaller and solve the high-dimension problem of text feature space. The semantic orientation and contribution of every single feature is taken into consideration. It makes the FS-BPSO-based feature selection method more suitable in discrete space and meets the characteristics of sentiment classification better.

D. Summarization

The particle swarm optimization selects some features, and then corresponding weights of those features are used to score the sentences. The score of the sentence is calculated by summing up the features weights corresponding to the bits containing ones and the features weights corresponding to the bits containing zeros are excluded from the scoring of the sentence. Based on the resulting scores for each sentence, the sentences are ranked in descending order. The top n of the sentences in the ranked list is selected as summary.

IV. RESULTS AND DISCUSSION

A. Dataset

The dataset consists of hotel reviews collected from TripAdvisor. It consists of ~98 MB of information. Extracted fields include date, review title and the full review in the following format:

Date<tab>Review title<tab>Full review

The proposed system uses the hotel review dataset and generates the complete summary.

B. Performance Evaluation

The evaluation metrics are important to identify the efficiency of the system. The performance of the system is calculated using the ROUGE-N metric. ROUGE is Recall Oriented Understudy for Gisting Evaluation used as performance metric to evaluate the quality of the summary generation. It measures the quality of summary by counting the overlapping N-grams between it and a set of reference summaries generated. ROUGE-N metric is given by the following formula,

$$\text{ROUGE}=\frac{S€\{\Sigma_{\text{ReferenceSummaries}}\}\Sigma_{\text{gram}\ n€S}\text{Count}_{\text{match}}(\text{gram}_n)}{S€\{\Sigma_{\text{ReferenceSummaries}}\}\Sigma_{\text{gram}\ n€S}\text{Count}(\text{gram}_n)}$$

Where n stands for the length of the n-gram and $\text{Count}_{\text{match}}(\text{gram}_n)$ is the maximum number of n-grams occurring in the candidate summary.

V. CONCLUSION AND FUTURE WORK

In this paper, a feature selection method based on binary particle swarm optimization is introduced and applied it to the sentiment classification domain. Then based on the feature selected text based summary is generated. Experimental results shows that proposed approach of summary generation is expected to improve the performance of summary by using multi-objective FBPSO when compared to probabilistic ranking approach. Future work focus on the use of binary particle swarm optimization for text summarization problem.

References

[1] B. Xue, M. Zhang, and W. N. Browne, “Single feature ranking and binary particle swarm optimisation based feature subset ranking for feature selection,” in Proc. 35th ACSC, vol. 122, Lecture Notes in Computer Science, Melbourne, Australia, 2012,pp.27–36.

[2] B. Xue, M. Zhang, and W. N. Browne, “New fitness functions in binary particle swarm optimisation for feature selection,” in Proc. IEEE CEC,2012,pp.1–8.

[3] Dim En Nyaung, Thin Lai Thein (2015), “Feature-Based Summarizing and Ranking from Customer Reviews”, International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol: 9, No.3.

[4] Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen (2006), “Opinion Extraction, Summarization and Tracking in News and Blog Corpora”. In Proceedings of AAAI.

[5] M. R. Sierra and C. A. C. Coello, “Improving PSO-based multi-objective optimization using crowding, mutation and epsilon-dominance,” in Proc. EMO, 2005, pp.505–519.

[6] Mohammed Salem Binwahlan, Naomie Salim, and Ladda Suanmali (2009), “Swarm Based Features Selection for Text Summarization”, IJCSNS International Journal of Computer Science and Network Security, Vol.9, No.1.

[7] Wang X, Yang J, Teng X, Xia W, Jensen R (2007), “Feature selection based on rough sets and particle swarm optimization”. Pattern Recognit Lett 28(4):459–471.

[8] Wang M, Cao D, Li L, Li S, Ji R (2014), “Microblog sentiment analysis based on cross-media bag-of-words model”. In: Proceedings of international conference on internet multimedia computing and service, p76.ACM.

[9] Zhou Z, Liu X, Li P, Shang L (2014) “Feature selection method with proportionate fitness based binary particle swarm optimization”. In: Simulated evolution and learning, pp 582 592. Springer, New York.