A Genetic Model with Semantic Analysis for Feedback Classification

Shanmuga Priya .A[1], Vijayakumar .P[2]

Research Scholar, Department of Computer Science, Sri Jayendra Saraswathy
Maha Vidyalaya College of Arts and Science, Coimbatore-5, India

Research Supervisor, Head, Department of Computer Application,
Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore-5, India

Abstract: ***An opinion, sentiment and topic detection system uses G-Tose algorithm and it's work is based on the semantic analysis and string co-occurrence rules. Compared with classical search and optimization, Genetic approaches are less stuck in local suboptimal regions of search space because they perform multiple searches for a solution. A G-Tose (GA) is a search based, self-learning algorithm that imitates the theory of natural evaluation based on selective screening of results based on fitness of purpose. The system evaluation results show precision of 90.25% respectively for feedback review process.***

Keywords: G-Tose=Genetic approach for Topic, opinion and semantic extraction.

1. Introduction

Genetic approach performs the discovery by extracting information from the original text documents. The first level is a preprocessing step that produces the initial population of GA from extracted information and training data. This training data's are stored in separate database for further evaluation and computation. It generates rule like representation for each document. Knowledge discovery based on genetic approach is the second level to produce the explanatory conclusions. Several generations are achieved, to get fixed user-defined values.

2. Problem Definition

With the increase of social sites and features on it provides more reliability and analytical tools against feedback summarizing issues which is not fully automated by the existing domain. For the purpose of effective and best organization of texts into relevant clusters is the main challenging task of data mining with weakly supervised and unsupervised data. The Motivations of choosing this opinion and topic from set of commands from online social website are that:

1. Social sites are the general domains which are having great scalability as well as more unlabeled data issues.
2. The proposed text mining for effective data clustering system can more easily be adapted to this type of social networks (because it contains many generic kinds of concepts or features)
3. It does not require a domain expert to understand the features and concepts involved.

Document clustering which automatically groups similar or related documents together has been used in practical applications to understand the contents and structures of documents in a better way.

3. Related Works

Researchers have experimented with several methods to solve the problem of opinion mining and topic detection using SVM, sentiwordnet, naivebayes etc.. But opinion mining is domain and context dependent problem .Support vector Vector machine is considered the best classification method (RuiXia, a 2011; Ziqiong, 2011; songho tan, 2008 and Rudy prabowo, 2009). But SVM seeks a decision from training datas. Thus SVM is effective only in the training set. Turney [1] used weakly supervised learning with similar information to predict the opinion by averaging out phrases related to sentiment within a document. White law et.al [9] used SVM to train different group features whereas Kennedy and Inkpen[10] leveraged two main sources i.e General Inquirer and choose the right word. Rudy prabowo(2009) describe the combination of rule based classification, supervised learning and machine learning into a new combined method.

The extension work was carried out by RuiXia (2011), which combines the output of several base classification models to form an integrated output. pang et.al[2] classified the movie reviews with supervised machine learning approaches and achieve the best result using SVM. Li and Zong[3] combined multiple single classifier trained on individual domains using SVM. Kaiquan XU, 2011 extract and visualize comparative relations between products from customer reviews for further design of new products. All the mentioned work above shares some similar limitations. They are;

1. They focused only on sentiment classification, without considering the topics of various domains in the text,
2. Most of the approaches favor supervised learning and required training datasets for classification.

4. Proposed System

In proposed system, Topic extraction and opinion mining is made with semi-supervised manner. Along with trained data sets, G-Tose algorithm is used to determine the solution. The new documents are preprocessed first which includes stop word removal process and stemming process. Then the new document is compared with samples and frequency values will be calculated. The threshold value is maintained to

obtain sentiment classification. Using G-Tose algorithm, new words are preprocessed and assigned in recursive steps.

4.1 Data Extraction

Text mining involves a series of activities to be performed in order to efficiently mine the information. These activities are: preprocessing text documents, analyzing those data by grouping into different categories and providing knowledge domain information is the final process of text mining.

Here text mining process has been implemented to find the category and opinion identification from the given dataset. Additionally this will help to filter text content which seems unwanted to display or eliminate in the user wall. The first step is extraction of data from the given dataset. The sentence which contains set of text will be extracted for the analysis. Identifying data's and splitting into terms is the major process.

In order to obtain all words that are used in a given input, a tokenization process is required, i.e. a text document is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized representation is then used for further processing. The set of different words obtained by merging all text documents of a collection is called the dictionary of a document collection.

4.2 Text Preprocessing

Many of the most frequently used words in English are useless in Information Retrieval (IR) and text mining. These words are called 'Stop words'. Stop-words, which are language-specific functional words, are frequent words that carry no information. Text preprocessing includes Word Disambiguation. This tries to resolve the ambiguity in the meaning of single words or phrases. Thus, instead of terms the specific meanings could be stored in the vector space representation.

The preprocessing process includes the stemming process, which eliminates unnecessary keys. The stemmer analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as cluster center. The Porter Stemmer has been used for English. All stemming algorithms can be roughly classified as affix removing, statistical and mixed. Affix removal stemmers apply set of transformation rules to each word, trying to cut off known prefixes or suffixes.

4.3 Term Selection

To further decrease the number of words that should be used also indexing or keyword selection algorithms can be used. In this case, only the selected keywords are used to describe the documents. A simple method for keyword selection is to extract keywords based on their entropy. Here the module represents different type of steps to extract term selection. First this extracts all texts and eliminating duplicate words in order to obtain unique terms. Pattern matching concepts has been applied in this module.

The main objective of document indexing is to increase the efficiency by extracting from the resulting document a selected set of terms to be used for indexing the document. Document indexing consists of choosing the appropriate set of keywords based on the whole corpus of documents, and assigning weights to those keywords for each particular document, thus transforming each document into a vector of keyword weights. The weight normally is related to the frequency of occurrence of the term in the document and the number of documents that use that term.

4.4 Semantic Orientation Scheme

The Semantic orientation approach to Sentiment analysis is, unsupervised learning" because it does not require prior training in order to mine the data. Instead, it measures how far a word is inclined towards positive and negative. The research in unsupervised sentiment classification makes use of the following processes.

- Synonym
- Antonym
- Hyponym

The semantic module also finds the co-clustering phase of text which is used to identify the meaning of the sentence. The sentence level clustering makes the proposed system in more effectively by providing exact opinion, sentiment and topic mining.

4.5 Genetic Model

The next module is the implementation of population generation which used by genetic approach for effective data classification. The document has been analyzed with prefixed labels. But the problem in the proposed method is text co-clustering. After the initial population is generated randomly, selection and variation function are executed in a loop until some termination criterion is reached. Each run of the loop is called a generation. The selection operator is intended to improve the average quality of the population by giving individuals of higher quality a higher. This process has been applied in the following way.

In genetic cross over and mutation concepts are applied. The system will analyze the word or set of words by synonym, antonym and hyponym terms and make the fitness test of those according to the label.

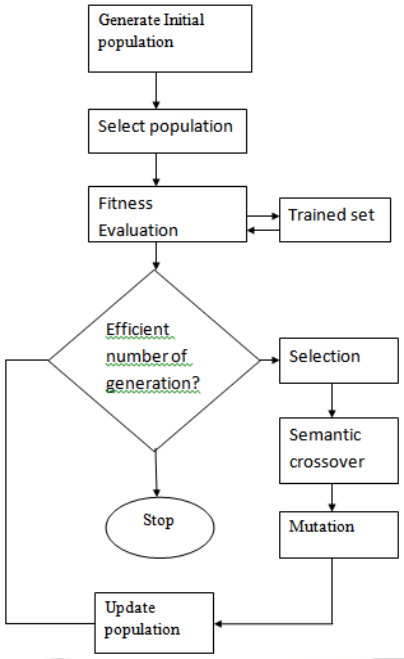

Figure 1: Work Flow Diagram of Genetic Model

4.5.1 Information Gain

The Genetic based Topic, Opinion and Semantic Extraction (G-TOSE) uses the information gain (IG) heuristic to weight the various sentiment attributes. These weights are then incorporated into the GA's initial population and crossover and mutation operators. For information gain (IG) Shanon entropy measure is used and is given as follows.

IG(C,A) = H(C) - H(C|A)

Where,

IG(C, A) = Information gain for feature A;

$H(C) = -\sum_{i=1}^{n} p(C = i) \log_2 p(C = i)$ = Entropy across sentiment classes C;

$H(C/A) = -\sum_{i=1}^{n} p(C = i/A) \log_2 p(C = i/A)$ =Specific feature conditional entropy;

n = Total number of sentiment classes;

If the number of positive and negative sentiment messages is equal, H(C) is 1.The information gain for each attribute A will vary along the range 0-1 with higher values indicating greater information gain. All features with an information gain greater than 0.0025 (i.e., IG(C, A)>0.0025) are selected. The use of such a threshold is consistent with prior work using IG for text feature selection.

4.5.2 Crossover

Crossover is the genetic operator that mixes two chromosomes together to form new offspring. Crossover occurs only with some probability (crossover probability). Chromosomes that are not subjected to crossover remain unmodified. The intuition behind crossover is the exploration of new solutions and exploitation of old solutions. GA's construct a better solution by mixing the good characteristic of chromosomes together. From the n solution strings in the population (simply n/2 pairs), certain adjacent string pairs are randomly selected for present crossover technique. In the standard GA, we use single point crossover by selecting a pair of strings and swapping substrings at a randomly. No adaptive or probabilistic crossover technique has been used for current experimentation.

4.5.3 Mutation

Each chromosome undergoes mutation with a probability. Crossover recombines two selected hypotheses and takes place with some probability, where both of them swap their elements at some random position of the hypotheses to produce new offspring. The coded data point obtained after the operation of selection on the sample pool of data points, is modified by tweaking one parameter to test its fitness factor and this continues for the entirely set of selected data points. The newly created generations of data-points are then tested using the selection operator again and the entirely process is restarted.

Algorithm: 1 G-TOSE algorithm

Input: document and words sets D and V; cluster numbers Kd and Kv; co clustered constraints M and C.
Initialize: Document and word cluster labels using k means.
Step1: Read the initial dataset
Step2: Preprocess the data using stemming algorithm, tokenizing also performed
Step3: Find unique T word and its frequency
Step4: Find cluster C
Step5: If (cluster/label found for the text T) then do step6
Step6: Add to the cluster
Step7: Else find semantic from data repository Sd. Find hyponym, synonym and do step 4
Step 8: Start co clustering process.
a. Read every named entity from cluster C
b. Co-cluster and find the frequency
Step9: Update cluster with conceptual clustering. Do cross over and mutation process
Step10: Return the groups with named entities.

In G_TOSE the new words are preprocessed and then words are assigned to cluster one by one in recursive steps. The new words are assigned to a cluster dynamically in run time without the need of re-clustering and also with automatic annotation of different key terms. As a result, the final step of clustering the proposed system will obtain the best evidence and provides effective topic or category of the set of words. For example a user uploaded a document with set of positive words, the system initially finds single clustering phase and annotate the labels. Finally it performs cross verification, mutation functions to confirm the opinion into a particular cluster.

5. Result Analysis

We have used c#.net for genetic algorithm application. The performance of the proposed algorithm is measured based on the detection speed, purity, reusability similarity and unlabeled data handling. Experiments were carried out to compare the performances of existing JST models and topic

clustering algorithms with Proposed G-TOSE Algorithm by varying the number of the documents. The Feedback information's are extracted from E-Commerce websites like www.amazon.com,www.mouthstop.com,www.consumerreview.com and www.epinions.com. The variation of clustering speed with the change in number of documents is also studied for these algorithms. Based on the theoretical analysis the below chart describes the time taken and the proposed system compared with the existing techniques in terms of time.

Table 1: Performance Comparison with Existing Model

Dataset	*Accuracy*	
	JST Model	*G-TOSE Model*
DS1	0.68	0.91
DS2	0.70	0.93
DS3	0.69	0.85
DS4	0.68	0.93

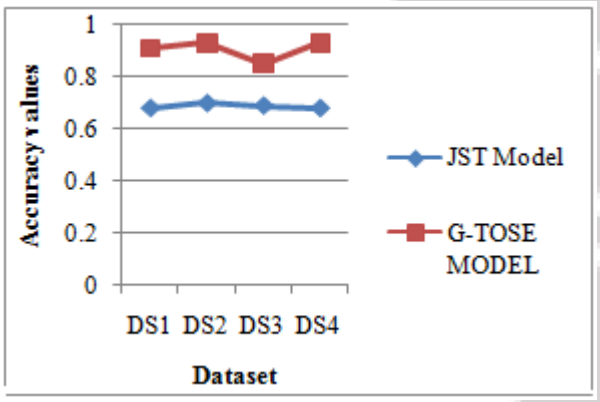

Figure 2: Performance of existing versus proposed Algorithm with respect to Time Accuracy

6. Conclusion

While most of the existing approaches to sentiment classification favor supervised learning, G-TOSE models target sentiment, opinion and topic detection simultaneously in a semi supervised fashion. The proposed system applies genetic approach for effective sub population creation based on available dataset. The investigation of unsupervised constraints is still preliminary. This will further investigate whether better text features that can be automatically derived by using natural language processing or information extraction tools. The future work may also interest in applying to other text analysis applications such as visual text summarization.

References

[1] P. D. Turney,"Thumps Up or Thumps Down? Semantic orientation applied to unsupervised classification of reviews," Proc. Assoc for computational Linguistics (ACL'01), pp.417-424, 2001.
[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumps up? Sentiment classification using a machine learning techniques," Proc. ACL Conf. Empirical Methods in Language processing (EMNLP), pp. 79-86, 2002.
[3] S.Li and C.Zong,"Multi-Domain Sentiment Classification," proc. Assoc. Computational linguistics-Human language Technology (ACL-HLT), pp. 257-260, 2008.
[4] Kushal Dave, Steve Lawrence, and David M. Pennoc, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews", In Proceedings of WWW, pages 519–528, 2003.
[5] Das and S. Bandyopadhyay, "Theme Detection an Exploration of Opinion Subjectivity", In Proceeding of Affective Computing & Intelligent Interaction (ACII 2009b).
[6] D. E. Goldberg "Genetic Algorithms in Search, Optimization and Machine Learning", Addison Wesley, New York, 1989.
[7] P. Foltz, W. Kintsch, and T. Landauer, "The Measurement of Textual Coherence with Latent Semantic Analysis," *Discourse Processes*, vol. 25, no. 2, 1998, pp. 259–284.
[8] Giuseppe Carenini, Raymond Ng, and Adam Pauls, "Multidocument summarization of evaluative text", In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pages 305–312, 2006.
[9] C. Whitelaw, N. garg and S. Argamon, "Using Appraisal Groups for sentiment Analysis", proc.14th ACM Int'l Conf. Information and Knowledge management(CIKM),PP. 625-631,2005.
[10] Kennedy and D. Inkpen,"Sentiment Classification of Movie Reviews Using Contextual Valence Shifters", Computational Intelligence, vol. 22, no. 2, pp. 110-125, 2006.
[11] M. Srinivas and L. M. Patnaik, "Adaptive probabilities of crossover and mutation in genetic algorithms", IEEE Transactions on Systems, Man and Cybernatics, 24(4):656–667, 1994.
[12] Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study," In the Proceedings of Recent Advances in Natural Language Processing (RANLP), 2005.
[13] Y. Seki, K. Eguchi, N. Kando, and M. Aono, "Multi-document summarization with subjectivity analysis at DUC 2005", in Proceedings of the Document Understanding Conference (DUC), 2005.

Author Profile

Shanmuga Priya A received her M.C.A degree from Nallamuthu Gounder Mahalingam College of Arts and Science in 2010.At present she is doing M. Phil in Sri Jayendra Maha Vidyalaya College of Arts and Science and her area of interest is Data Mining.

P. Vijayakumar M.C.A., M.Phil., working as head with twelve years of experience in the department of computer application, Sri Jayendra Maha Vidyalaya College of Arts and Science, Coimbatore-5. His area of interest is Data Mining.