ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.9

Feature Clustering using Simulated Annealing Technique

Rupali T. Jamdhade.[1], Mr. K. P. Gaikwad[2]
PG Student[1], Professor[2]
Department of Computer Science and Engineering
MIT College, Aurangabad, India

Abstract:
Text mining techniques helps users to find useful information from a bulky quantity of digital text documents on a Web or database engines. Therefore it is important that a good text mining model should retrieve the information that meets users' needs within a relatively efficient time frame. Traditional Information Retrieval (IR) has the similar objective of automatically retrieving relevant documents as many as possible while filtering out non-relevant ones at the same time. To assure the quality of discovered relevance feature in text documents for describing the user preferences is extremely important and demanding task because of big scale of terms and data patterns. The popular existing text mining and classification methods adopt term based approaches which suffer from the problem of polysemy and synonymy. Assumption is that the pattern based methods performs superior than term based ones. A pattern discovery approach for text mining discovers frequent sequential pattern and closed sequential patterns in text documents for identifying the most information contents of the documents and extract valuable features for text mining. It also classifies extracted terms into three categories: positive terms, general terms, and negative terms. In this project new clustering algorithm is implemented to cluster terms into three categories. The SA algorithm starts from a random initial configuration. Partitioning of the data is based on the minimum squared distance criterion.

Keywords: text mining, pattern discovery, clustering.

I. INTRODUCTION

Text mining is the process of discovering valuable knowledge in text documents. It is difficult problem to find correct knowledge in text documents to help out users to discover what they want. Search engine can returns millions of document for single query. In text classification relevant feature focuses on identifying relevant information according to user need, According to the IR mainly there are four methods used. Term Based Method (TBM), Phrase Based Method (PBM), Concept Based Method (CBM), Pattern Taxonomy Method (PTM). In term based method document is analyzed on the basis of term and has reward of efficient computational performance as well as mature theories for term weighting. Term based method suffer from two issues polysemy and synonymy. In pattern based method documents are analyzed on pattern basis. Pattern mining has been extensively studied in data mining communities for many years. A variety of efficient algorithms such as Apriori-like algorithms, PrefixSpan, FP-tree, SPADE, SLPMiner and GST [4], [5],[6],[7] have been proposed. Patterns can be discovered by data mining techniques like frequent item set mining, sequential pattern mining and closed pattern mining [2]. Patterns can be represented into taxonomy with is-a relation. Pattern mining has been widely studied in data mining. Patterns are discovered by data mining techniques. Techniques are association rule mining, frequent item set mining, sequential pattern mining and closed pattern mining [2]. There are two challenging issues in using pattern mining techniques for finding relevance features in both relevant and irrelevant documents low-frequency and misinterpretation problems for text mining[5]. Two challenges using the pattern matching techniques are low support problem and misinterpretation. The aim of relevance feedback is to find useful information available in feedback documents including both positive and negative documents. Due to large amount of terms, patterns and noise there is difficulty in discovery of relevance features in text document for describing user's need. Pattern-based model used for the representation of text documents**.** In PTM, documents split into set of paragraphs and each paragraph consists of set of words. Data mining techniques are applied to find frequent patterns and generate pattern taxonomies. Pattern taxonomy is a tree-like structure that shows the relationship between extracted [3]. PTM lacks in use of negative document. Assumption is that negative document is useful to search accurate information.RFD model uses both high level feature and low level features to overcome the low frequency patterns To select negative documents (so-called offenders) that are closed to relevant document, it also introduces a method. Selected negative documents are used to calculate specificity of each term. Specificity function is introduced to know what the topic is focused on. Then low level terms are clustered into three categories using Fclustering and SA(simulated annealing).

II. SYSTEM DEVELOPMENT

The proposed system clustered the discovered terms (low level feature) into three categories positive terms, negative terms and general terms based on specificity. Features are discovered from both positive and negative documents. Proposed approach evaluates weights of terms according to both their specificity and their distributions in the patterns include both positive and negative patterns. The traditional feature selection methods are not efficient for selecting text features for solving relevance. The efficient way of feature selection for relevance is based on a feature weighting function. A feature weighting function indicates the degree of information represented by the feature occurrences in a document and reflects the relevance of the feature.

A .Text Pre-Processing: Data preprocessing reduces the size of the input text documents significantly. It involves activities like specific stop word elimination and stemming. Stop-words are functional words which occur frequently in the language of the text (for example a, the, an, of etc. in English language), so that they are not useful for classification.

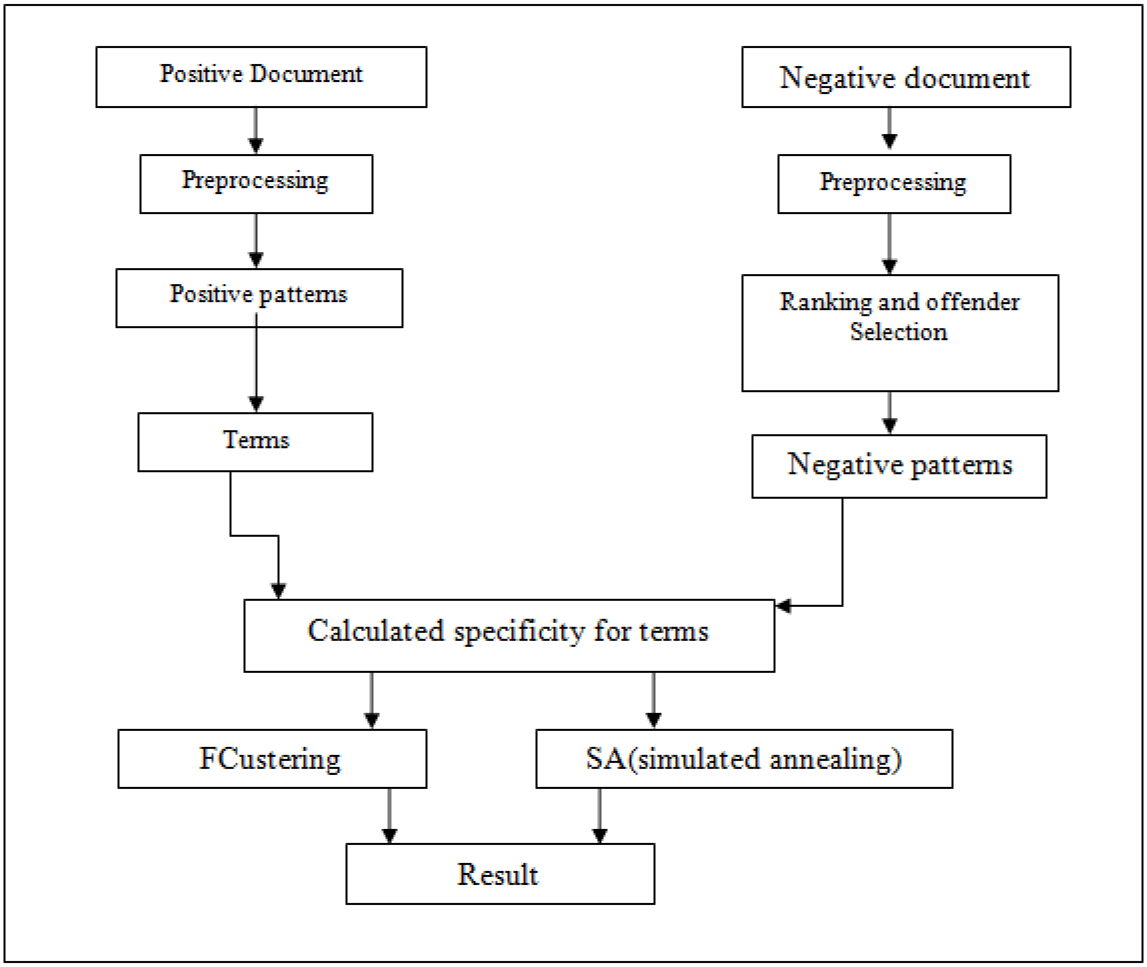


Figure.2. Block of Proposed System.

Once the data is pre-process then it will be the collection of the words. [1]The purpose of stemming method is to remove various suffixes, to reduce the number of words, to save time and memory space.

B. Pattern discovery

High level features (patterns) are extracted from positive document set. Document is split into a set of paragraphs. Each paragraph consist of set of words consider as transaction [4]. Sequential pattern mining algorithm (Apriori) extracts large number of patterns. To reduce number of patterns threshold value (min_sup) is used.

Definitions:

The absolute support of P is the number of occurrences of P in d.$Supp_a(P) = |\{S \mid S \in d, P \sqsubset S\}|$

The absolute support of P is the number of occurrences of P in d.

$Suppr(P) = Suppa(P) \ /|dp|$

Where | dp| is number of paragraphs in document d.

A frequent sequential pattern p_1 is a closed pattern if there is no frequent sequential pattern p_2 such that $p_1 \sqsubset p_2$ and $Suppa(p_1) = Supp_a(p_2)$.

C. Weight calculation:

Let $SP_1, SP_2, \ldots, SP_n$ be the sets of discovered closed sequential patterns for all documents $d_i \in D^+ (i = 1,\ldots\ldots,n)$, where $n = |D^+|$. Weight of each term t is calculated using equation (1).

$$weight_{(t, D+)} = \sum_{i=1}^{n} \frac{|\{S \mid SPi \in d, t \in P\}|}{\sum_{P \in SPi} |P|}$$

D. Offender selection: To improve the performance negative documents are also used. But not all documents are useful, so what to select negative document which are closed to positive documents [4]. For that offender selection method is used rank the negative document using rank equation. Select the *n/2* document from ranked negative document set. Where *n* is number of positive document. Once we select the top-K negative documents, the set of negative document will be reduced to include only K offenders (negative documents).

$Rank\ (d) = \sum_{t \in T} weight(t)\tau(t,d)$

Where τ(t, d) = 1 if t Є d; otherwise τ(t, d) = 0.

E. Calculating term's specificity:

Given a term t belongs to T, its *coverage+* is the set of positive documents that contain t, and its *coverage*$^{-}$ is the set of negative documents that contain t.

$Spe\ (t) = \frac{|coverag\ e^{+}\ (t)|-|coverage\ ^{-}(t)|}{n}.$

Spe (t) > 0 means that term t is used more frequently in positive documents than in negative documents. We present the following classification rules for determining the general terms G, the positive terms T^{+}, and the negative terms T^{-}.

F. Recalculating the weight:

Following principles: increment the weights of the positive terms, decline the weights of the negative terms, and do not update the weights of the general terms.

G. Algorithms:

FClustering Algorithm:

Describes the process of feature clustering [5], where DP+ is the set of discovered patterns of D^{+} and DP is the set of discovered patterns of D-.

Step:

1. Initialize the three categories.
2. All terms that are not the elements of positive patterns are assigned to category T^{-}.
3. For the remaining m terms, each is viewed as a single cluster in the beginning.
4. It also sorts these clusters in C_i based on their min_spe values.
5. Iterative process of merging clusters until there are three clusters left.
6. In the last step, it chooses the first cluster as T^{+}, the second cluster as G and the last cluster as a part of T^{-}.

Simulated annealing:

Is a method for finding a good (not necessarily perfect) solution.[9] It is based on the minimum squared distance criterion. An optimization algorithm searches for the best solution by generating a random initial solution and "exploring" the area nearby. If a neighboring solution is better than the current one, then it moves to it. If not, then the algorithm stays put.

This technique is easy to implement.

1. First, generate a random solution
2. Calculate its cost using some cost function.
3. Generate a random neighboring solution
4. Calculate the new solution's cost.
5. Compare chosen solution with new solution.
a. If new solution < old solution
 Then move to the new solution
b. If new solution > old solution
Then move to the old solution
6. Repeat steps 3-5 above until an adequate solution is found.

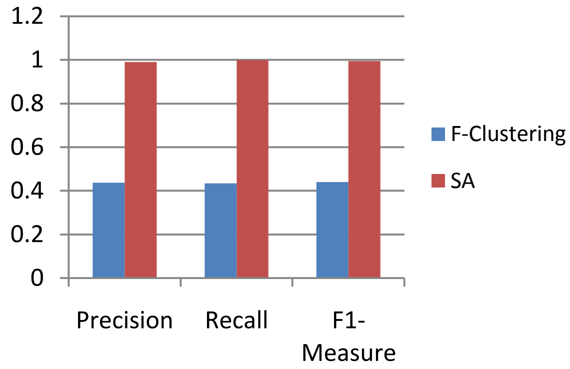
III. EXPRIMENT RESULT

To evaluate this work Reuters21578 dataset is use for evaluation. Reuters corpus volume is an archive consist of news stories, which are manually categorized. Each category consist different number of documents related to category name. Performance measures used are precision recall and f1 measure. Fclustering and SA algorithm compare on the basis of this three measures.

Table.1. Comparison of parameter using Fclustering and simulated annealing.

Performance measure	Fclustering	Simulated Annealing
Precision	0.4369	0.9900
Recall	0.4339	0.9984
F1 measure	0.4400	0.9953

GRAPH.I. Performance Comparison using F -Clustering and SA



IV. CONCLUSION

The proposed method used to discover features (terms). With consideration their appearance in patterns and specificity function used to classify this terms. To improve the performance offender section method is introduced. Specificity function is reasonable and used correctly weight the terms. The term classification can be effectively approximated by a feature clustering method. The clustering method is effective and the results show that the proposed specificity function is adequate. The paper shows that the use of negative feedback is important for improving the performance of relevance feature discovery models. It provides a hopeful methodology for developing effective text mining models for relevance feature discovery based on both positive and negative feedback. In this FClusreing and SA algorithms are used for feature clustering. SA is based on the minimum squared distance criterion and gives good solution.

V. REFERENCES

[1]. Text Mining Methods and Techniques International Journal of Computer Applications (0975-8887) January 2014.

[2]. S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic pattern taxonomy extraction for web mining,

[3]. S.-T. Wu, Y. Li, and Y. Xu,, “Deploying approaches for pattern refinement in text mining,” in Proc. IEEE Conf. Data Mining, 2006

[4]. *Y.* Li, A. Algarni, and N. Zhong, “Mining positive and negative patterns for relevance feature discovery,” in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2010, pp. 753–762.

[5]. Relevance Feature Discovery for Text Mining, IEEE Transactions June 2015.

[6]. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases, pages 487{499. Morgan Kaufmann Publishers Inc., 1994.

[7]. R. Agrawal and R. Srikant. Mining sequential patterns. In Data Engineer- ing, 1995. Proceedings of the Eleventh International Conference on, pages 3-14. IEEE, 1995.

[8]. Prabha S , Shanmugapriya S , Duraiswamy K, A Survey on Closed Frequent Pattern Mining**.** International Journal of Computer Applications (0975 – 8887) Volume 63– No.14, February 2013.

[9]. Ujjwal Maulik, Sanghamitra Bandyopadhyay. Performance Evaluation of Some Clustering Algorithms and Validity Indices. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 12, DECEMBER 2002.

[10]. Manoj Ganjir, Jharna Chopra**.** Combining Apriori and FP Growth algorithms with Simulated Annealing for Optimized association Rule Mining. International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12, December 2015.