International Journal of Soft Computing 9 (3): 178-182, 2014
ISSN: 1816-9503

Classifier Using Conceptual Granulation and Equal Partition Approach

[1]D. Malathi and [2]S. Valarmathy
[1]Department of CA, [2]Department of ECE,
Bannari Amman Institute of Technology, Tamil Nadu, India

Abstract: This study presents a systematic approach for the classification of large corpus based on concept granulation and equal partition approach. The proposed research has three main processes which are the preprocessing treatments to text documents, feature extraction and finally the classification. The proposed approach is concentrated in the feature extraction phase. Almost bird eye view like approach is the feature extraction method. So, the proposed research concept granulation and equal partition approach has been named as Immune Term (TIM) which finds the immunized terms from the information system. At first, documents are preprocessed from text to numerical form, i.e., word frequency is calculated for each document. Second, sets of features are extracted using TIM. In the third step, the TIM treated feature is introduced to Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI) for global set extraction or dimension reduction. Finally, Naive Bayes (NB) and Support Vector Machine (SVM) are used to classify the documents. The proposed research seems to be fruitful when compared to the conventional word frequency approach.

Key words: Domain classifier, concept granulation, equal partition, global set, eye

INTRODUCTION

Information retrieval for knowledge extraction is the food served from World Wide Web today. The research communities are setting mile stones with new approaches for fruitfulness of truth accessed from internet. The research researchers present here is a feature extraction by spreading word granules and combining by micro averaging between local spreads. By this method the concrete terminological words for topics are highlighted. Further, this approach has been realized to be better performer when compared to the conventional probabilistic and statistical approaches.

Text mining: Text mining is the discovery by computer of new, previously unknown information by automatically extracting information from different written resources (Feldman and Sanger, 2007). Text learning technique is nothing but extraction of information from words. A role of text mining is linking together the extracted information to form new hypotheses which has to be explored by the Existing Means of algorithms and applications.

Preprocessing: The initial preprocessing step in text mining is tokenization process, i.e., a text document is split into a bag of words by removing all punctuation and non text characters. This tokenized representation is further reduced into set of words by the removal of stop words like the, are, is, of and so on. These stop words are removed because they do not support the proposed method of inferring knowledge. Further, the size of the words is trimmed by stemming process. The tokenization process is common for all the text mining research.

LITERATURE REVIEW

Domain classifier in text mining is nothing but topic modeling. In real world, the sequential patterns in natural language usually appear without explicit boundaries but with the variations of temporal topics (Chien and Chueh, 2012). When compared to Vector Space Model (VSM), Topic Model (TM) represents documents in much lesser dimensional space (Hofmann, 1999).

The curse of dimensionality and its reduction is studied and surveyed (Malathi and Valarmathy, 2011). Further statistical composition of patterns from text can be learned from dimensionality reduction methodologies. To follow up this survey, VSM is inclined with Singular Value Decomposition (SVD) and classified using Support Vector Machine (SVM) (Malathi and Valarmathy, 2012).

The feature extraction phase is found to be the primary cause for knowledge associate for the research on text categorization. Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) is known as a latent topic document model which discovers the semantic structure in training documents. It calculates the word probability by:

Corresponding Author: D. Malathi, Department of CA, Bannari Amman Institute of Technology, Tamil Nadu, India

$$p(w|d)=\sum_{c=1}^{n} p(w|c)p(c|d)$$

where, the equation denotes the distribution of word w in topic c and the distribution of topic in a document d. LDA extends PLSA (Blei *et al*., 2003) model in the Document Segmentation Model (Chien and Chueh, 2012). It consists of a column vector of dirichlet distribution with parameter α, topic word probabilities β and topic sequence θ. The Document Segmentation Model with dLDA (Brants *et al*., 2002) is formed by sequence of block bk. The distribution of word in each block and topic distribution are determined. Similarity between each block bk and bk+1 is computed using cosine equation:

$$s(b_k, b_{k+1}) = \frac{\sum p(w|b)p(w|b_{k+1})}{\sqrt{\sum p(w|b_k)^2}\sqrt{\sum p(w|b_{k+1})^2}}$$

The Naive Bayes classifier (Lewis, 1992), a simple Bayesian Classification algorithm is an effective basic approach for text categorization. This has been taken up for testing the TIM algorithm with the conventional tf-idf:

$$P(c|d)\alpha P(c).P(d|c)$$

By Naive Bayes way, the words in the documents are conditionally independent and is given by:

$$P(c|d)\alpha P(c)\prod_{w\in d} P(w|c)$$

The SVM (Tong and Koller, 2001) classifier is well suited for text categorization and acknowledges the sparse data. SVM fixed data in the hyper plane. SVM is best suited for the binary classification. It separates dataset into classes with the help of Kernel functions. The Kernel functions depends of the complexity of the data. They are linear and polynomial. The details of SVM is clearly studied by Manning *et al*. (2008).

SYSTEM STUDY

An information table (Pawlak, 1982) represents complete information and knowledge, i.e., the objects are processed, observed or measured by the finite number of attributes.

Definition 1 (Information system): An information system is a pair S = (U, A) with every a∈A, a set of its values V_a is associated. It is expressed as:

$$S = (U, A, \{V_a \mid a \in A\}, \{I_a : U \to V_a \mid a \in A\})$$

Where:

U = Universal finite non empty set of objects
A = Finite non empty set of attributes
V_a = Domain of attributes a∈A
I_a: U→V_a = Information function that maps an object in U to V_a

The information system has to be presented in such a way that for any related classification problem, a correct decision can be derived.

Definition 2 (Document information): Document information is defined by information system as a pair P = (D, T), set of its values V_τ is associated where τ∈T. It is expressed as:

$$P=(D,T,\{V_\tau \mid \tau \in T\},\{I_\tau : D \to V_\tau \mid \tau \in T\})$$

Where:

D = Document collection
T = Finite terms
V_τ = Domain of terms
I_τ: D→Vτ = Information function that maps terms in D toV_τ

DATASETS

Movie review dataset: The movie review dataset, polarity dataset v0.9 with 700 positive and 700 negative reviews is used. Using movie reviews as data, the problem of classifying documents using standard machine learning techniques definitively outperform human-produced baselines processed reviews (Pang *et al*., 2002). The training cases are chosen randomly from each classes about 100 documents each. Which means about 200 cases is considered for training.

Reuters-21578 data set: The reuters-21578* (*http://www.daviddlewis.com/resources/testcollections/reuters21578) data set collection provides a classification task with challenging properties. There are multiple categories, the categories are overlapping and non-exhaustive and there are relationships among the categories. There are interesting possibilities for the use of domain knowledge. There are many possible feature sets that can be extracted from the text and most plausible feature/example matrices are large and sparse.

PROPOSED APPROACH

The proposed approach TIM (Immune Term) has been inspired from segment based approach (Chien and

Chueh, 2012; Brants *et al*., 2002; Boley *et al*., 1999). TIM is a local approximation approach in which each term is treated as a granule in each document. Each document is segmented or partitioned into a constant k. If a term t is distributed throughout the partitions and is satisfies a threshold then it is included in the global set as t':

$$t' = \frac{\sum p\left(\frac{\Gamma}{\kappa}\right)}{\kappa}$$

Where:

$$\kappa = \sum_{l=1}^{k} (t_l \,|\, \Gamma_l > 0)$$

The algorithms TIM and conceptgranule are designed is such a way to check whether a term is immunized, i.e., t<< t' (Dimension reduction achieved):

```
Algorithm: TIM (Corpus)
Input: 'n' No. of training Documents D
       No. of document Partitions 'k'
       No. of local terms 't'
       gm[][]-granule matrix with 't' terms
       im[][]-immune matrix with 't' terms
Output: Reduced Documents with
        local immune terms t'
1.   get value for k
2.   for i = 1 to n              // training documents
3.     Preprocess each document by trimming
                          and stop word removal
4.     Initialize a granule matrix gm[][] with t terms
5.     z = 0
6.     for j = 1 to t              // terms
7.       for 1 = 1 to k          // partitions

8.                 Γ[l] = gm[j][l]

9.                 If conceptgranule(Γ, k)     then

10.                 im[i][++z] = conceptgranule (Γ, k)
```

Reduction of bag of words to concept granule is the feature extraction adopted in the above said algorithm. If U is the Universal set, i.e., bag of words of document, then g, the set of concept granule set is obtained by calculating the frequency of each word in each partition.

Dj is a matrix with k × i matrix with vector gik where 'i' represents term or words 'k' represents partition of a document. ti'j is a single vector or a set of words in a document, where i' represents immune word, i.e., each word is treated for its importance in the whole document by the algorithm. When a word is able to withstand the treatment then that is said to be immune word and i represents document.

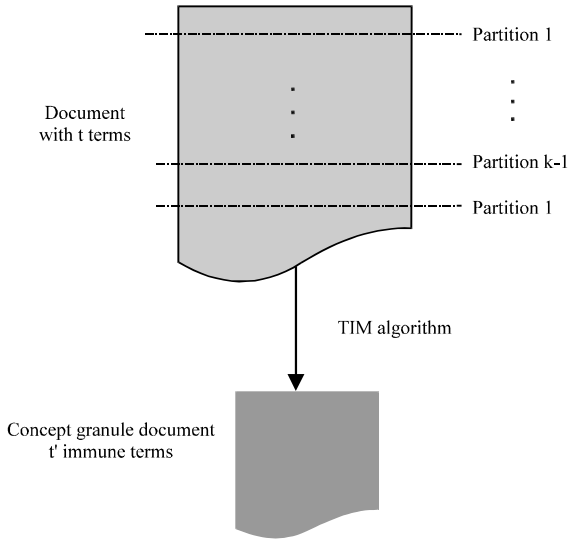

Fig. 1: TIM representation

```
Algorithm: Conceptgranule(Γ, k)
Input:   Γ term array with   k-terms
Output: Non zero value for immune terms
            or zero for non immune terms
1.    κ = 0
2.    for l = 1 to k                        // partitions

3.          if   (Γ[l])   then

4.                  ++   κ
5.          if κ≤3×k/4   then     //a threshold
6.                for l = 1 to k        // partitions

7.                   t' = t'+Γ_l

8.                  r = t'/κ
9.                  return (r)
                //include in term frequency matrix
10    else
11.         return (0)
```

The TIM approach is supposed to studyas mentioned in the Fig. 1

EXPERIMENT AND RESULTS

The experiment is taken with the reuters and movie review data sets. The connection between words and the respective topic are taken into consideration. At first the training to the topics such as "wheat", "trade", "ship", etc. are taken up in reuters and star level such as "1", "1.5", "2", etc. are taken up in movies review datasets. This is done with random selection of documents with the respective topic. Further, testing is done for each of the trained topic. The result is studied by micro averaging the topics and presented in the Table 1.

The TIM algorithm which has been proposed, shown positive play in reduction (using PCA dimension and LSI)

Table 1: Micro average of PCA and LSI reduced NB and SVM classified reuters and movie review datasets

		tf-idf		TIM	
Datasets	Classifiers	PCA	LSI	PCA	LSI
Reuters	NB	0.785	0.791	0.912	0.921
	SVM	0.824	0.892	0.939	0.932
Movie review	NB	0.722	0.706	0.811	0.792
	SVM	0.788	0.745	0.852	0.874

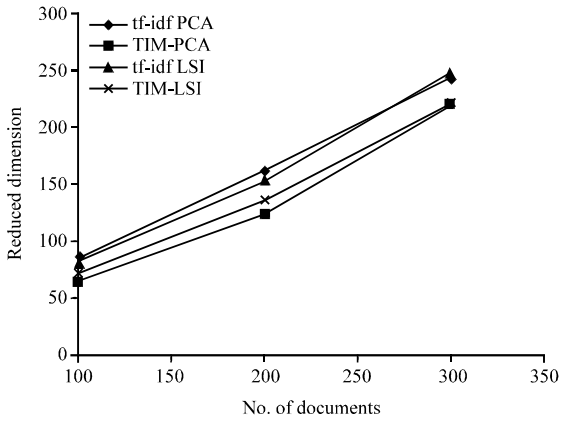

Fig. 2: Dimension reduction of tf-idf and TIM preprocessing

and classification (using NB and SVM). This could be comparably seen in the dimension reduction scheme represented in the Fig. 2.

On the reuters the topical word support is much when compared to the movie reviews. This is reflected in the results of the classification. Reuters could be classified at higher end whereas the movie review could be able to respond to level less. The main concentration is given to the preprocessing level of documents. TIM algorithm is introduced as next phase to tf-idf. The results shows much better improvement in dimension reduction and classification.

One advantage of using TIM approach is, documents are treated much in lower level, than the inclusion of ontology or parts of speech tagging. Because of this, much of the time is reduced in preprocessing level.

One difficulty of using TIM approach is that each of the training documents is partitioned into k segments in the training level. Because of this the training time is fat but testing time is much reduced by looking only for the topical words.

CONCLUSION

TIM algorithm developed with the inspiration of concept granulation and equal partition based approach has found to be yielding better performance result with the support of PCA, LSI, NB and SVM approaches. The magnitude of terms is playing the major role of the algorithm.

Though PCA, LSI, NB and SVM are highly repeated techniques, they seem to be much beneficial in understanding the proposed research. The TIM-PCA and TIM-LSI have marked computationally changes in the dimension. In spite of this one important point is noted. The datasets play the major role in classification. The reuters dataset is topic oriented and has been classified with more than ten topics. The movie review is not topically distributed. It is proposed for sentimental classification, i.e., positive and negative sentiments. The major role is played by the adjective and adverb terminologies, rather than the topical words.

NB and SVM is showing low performance in the movie review dataset than the reuters dataset. It has been understood that the preprocessing of documents much in the lower level or local optimization is very essential for unstructured data.

REFERENCES

Blei, D.M., A.Y. Ng and M.I. Jordan, 2003. Latent dirichlet allocation. J. Mach. Learn. Res., 3: 993-1022.

Boley, D., M. Gini, R. Gross, E.H.S. Han and K. Hastings *et al*., 1999. Partitioning-based clustering for web document categorization. Decis. Support Syst., 27: 329-341.

Brants, T., F. Chen and I. Tsochantaridis, 2002. Topic-based document segmentation with probabilistic latent semantic analysis. Proceedings of the Eleventh International Conference on Information and Knowledge Management, November 4-9, 2002, McLean, Virginia, USA., pp: 211-218.

Chien, J.T. and C.H. Chueh, 2012. Topic-based hierarchical segmentation. IEEE Trans. Audio Speech Lang. Process., 20: 55-66.

Feldman, R. and J. Sanger, 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, ISBN: 9780521836579, Pages: 410.

Hofmann, T., 1999. Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA., USA., pp: 50-57.

Lewis, D.D., 1992. Representation and learning in information retrieval. Ph.D. Thesis, University of Massachusetts Amherst, MA, USA.

Malathi, D. and S. Valarmathy, 2011. A comprehensive survey on dimension reduction techniques for concept extraction from a large corpus. Int. J. Comput. Inform. Syst., 3: 1-6.

Malathi, D. and S. Valarmathy, 2012. Conceptually Co-occurring words included as feature selection in text document classification using SVD and SVM. Int. J. Adv. Res. Comput. Sci., 3: 145-148.

Manning, C.D., P. Raghavan and H. Schutze, 2008. An Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK., ISBN-13: 9780521865715, pp: 482.

Pang, B., L. Lee and S. Vaithyanathan, 2002. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods on Natural Language Processing, July 6-7, 2002, Philadelphia PA., USA., pp: 79-86.

Pawlak, Z., 1982. Rough sets. Int. J. Comput. Inform. Sci., 11: 341-356.

Tong, S. and D. Koller, 2001. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res., 2: 45-66.