Syntax and Semantics based Efficient Text Classification Framework

Suganya . S
PG Scholar
Sri Ramakrishna Engg College
Coimbatore,Tamil Nadu,India

Gomathi . C
PG Scholar
Sri Ramakrishna Engg College
Coimbatore,Tamil Nadu,India

ManoChitra .S
PG Scholar
Sri Ramakrishna Engg College
Coimbatore,Tamil Nadu,India

ABSTRACT
This system proposes an efficient text classification approach which is based on multi – layer SVM-NN text classification and two-level representation model. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, which can be too expensive. This system proposes two-level representation model to represent text data, one is for representing syntactic information using tf-idf value and the other is for semantic information using Wikipedia. Further, a multi-layer text classification framework is designed to make use of the semantic and syntactic information. The proposed framework contains three SVM-NN classifiers in which two classifiers are applied on syntactic level and semantic level in parallel. The outputs of these two classifiers will be combined and given as input to the third classifier, so that the final results can be obtained.Experimental results on benchmark data sets like 20Newsgroups and Reuters-21578 have shown that the proposed model improves the text classification performance.

Keywords
Wikipedia, Semantics, Text classification, Text representation, Multi-layer classification, SVM.

1. INTRODUCTION
Text (or Document) classification is an active research area of text mining. Text classification is the task of automatically sorting a set of documents into categories from a predefined set. Mostly text documents include letters, newspapers, articles, blogs, technical reports, proceedings, and journal papers, etc. The task of text categorization is to build a classifier based on some labeled documents and to classify the unlabeled documents into the prespecified categories. To structure text data, Bag of Words (BOW) model i.e. term-based Vector Space Model (VSM) is most widely used. Term-based VSM has been widely applied to text categorization due to its simplicity and good performance. However it does not consider the semantic relatedness between words. Different approaches have been made to introduce semantic information to text representation with the aid of background knowledge bases such as WordNet and ODP2. A good solution is utilizing both term and concept information. Even though such integrated representation models were experimentally shown to improve the performance of text categorization, how to sufficiently apply syntactic and semantic information is still an open problem. In order to make use of syntactic and semantic information, the proposed system uses a two-level representation model. It represents document in a two-level vector space containing syntactic (term) and semantic (related concept) information respectively. The syntactic level represents each document as a term vector and the component records tf-idf value of each term. The semantic level represents document with Wikipedia concepts related to terms in syntactic level. A context-based method is adopted to identify the appropriate Wikipedia concepts.

Based on two-level representation model, a multi-layer SVM-NN (Support Vector Machine - Nearest Neighbor) text classification framework is used to analyse text data in a way of layer-by-layer. Multi-layer classification framework includes three classifiers. Among them, two SVM-NN classifiers are applied on syntactic and semantic level independently. Based on the output of these two classifiers, each document is represented as two compressed vectors. The combined vector from the above two compressed vectors will be input to the third SVM-NN classifier to obtain the final results. This framework effectively keeps the primary information and reduces the influence of noise by compressing the original information, so that the proposed framework guarantees the quality of classification results.

2. RELATED WORK
Hotho et al., 2003[1], proposed classification system by taking the synonyms in WordNet of each term as the related concepts. Although empirical results have shown this method was efficient in some cases, wordnet is manually built and its coverage is far too restricted.Gabrilovich and Markovitch.,2005[2], used machine learning techniques to map document to the most relevant concepts in ODP or Wikipedia by comparing the textual overlap between each document and article. However, its feature generation procedure requires high processing efforts, because each document needs to be scanned multiple times. Besides, it produced too many Wikipedia concepts for each document and filtering step further increases the processing time.

Gabrilovich and Shaul Markovitch.,2006[3], proposed a method, titled Explicit Semantic Analysis (ESA), for grained semantic interpretation of unrestricted natural language texts. It represents meaning in a high-dimensional space of concepts derived from Wikipedia, the largest encyclopedia in existence. It represents the meaning of any text in terms of Wikipedia-based concepts.

Somnath Banerjee, Krishnan Ramanathan and Ajay Gupta.,2007[4], clustered similar items in the feed reader to make the information more manageable for a user. This method improved the accuracy of clustering short texts by enriching their representation with additional features from Wikipedia. Empirical results indicate that this enriched representation of text items can substantially improve the clustering accuracy when compared to the conventional bag of words representation.Chen and Hu et al. ,2008 [5],constructed an informative thesaurus from Wikipedia so that the synonymy, polysemy, hyponymy, and associative relations between concepts can be explicitly derived. But they rely on an exact phrase matching strategy while this strategy is limited by the terms appearing in the documents and the

coverage of Wikipedia concepts or article titles. Wang et al. ,2008[6], adopted a kernel method to enrich document representation matrix, which replaced concept similarity matrix for the term similarity matrix. Concept similarity matrix was measured by taking account of synonyms, hyponyms and associative concepts in Wikipedia. However, These methods do not use the contextual semantic within the document.

Ming-Wei Chang, Lev Ratinov, Dan Roth and Vivek Srikumar [7], 2008, proposed learning protocol that uses world knowledge to induce classifiers without the need for any labeled data. Like humans, a dataless classifier interprets a string of words as a set of semantic concepts.

3. PROPOSED SYSTEM

The proposed efficient text classification framework integrates the two–level representation model with multi – layer SVM-NN text classification. Efficient text classification with increased accuracy and reduced noise problems are proposed in the system.

3.1 Two – level representation model

In this Section, the Two-level Representation Model that represents syntactic information and semantic information with two levels are proposed. Term-based representation and tf-idf weighting scheme are used in syntactic level to record the syntactic information. Semantic level consists of Wikipedia concepts related to the terms in the syntactic level. These two levels are connected via the semantic correlation between terms and their relevant concepts.

3.1.1 Constructing term based representation model

Term based Representation Model represents syntactic information for Term-based representation using tf-idf weighting scheme. Term frequency (TF) is essentially a percentage denoting the number of times a word appears in a document. It is mathematically expressed as

$$C / T \quad (1)$$

Where C is the number of times a word appears in a document, T is the total number of words in the same document. Inverse document frequency (IDF) takes into account that many words occur many times in many documents. It is mathematically expressed as

$$IDF_k = \log (n / n_i) \quad (2)$$

Where n denotes number of documents, n_i denotes number of documents in which the term k occurs .Term Weight is a measure used to calculate weight of term which is scalar product of term frequency and inverse document frequency mathematically represented as

$$W=tf*idf \quad (3)$$

3.1.2 Constructing concept based representation model

Semantic level consists of Wikipedia concepts related to the terms in the syntactic level . These two levels are connected via the semantic correlation between terms and their relevant concepts.The semantic relatedness between term and its candidate concepts in a given document is computed in equation 4 as

$$Rel(t,c_i|d_j)=1/|T|-1\sum_{t1\varepsilon T\&t1\neq t}1/|CSl|\sum_{ck\varepsilon csl}SIM(c_i,c_k) \quad (4)$$

Where T is the term set of the jth document d_j , t_l is a term in d_j except for t,cs_l is the candidate concept set related to $term t_l$.$SIM(c_i, c_k)$ is the semantic relatedness between two concepts, which is calculated with the Wikipedia hyperlinks as

$$SIM(c_i,c_k)=\log(\max(|A|,|B|)\log(|AnB|)/\log(|W|)-\log(\min(|A|,|B|)) \quad (5)$$

Where A and B are the sets of all articles that link to concepts c_i and c_k ,W is the set of all articles in Wikipedia. The concepts with highest relatedness will be used to properly build the concept vector in semantic level. Based on Rel(t, c_i | d_j) and term's weight $w(t_k,d_j)$, the concept's weight is defined as their weighted sum as follows

$$W(c_i,d_j) = \sum_{t_k \varepsilon T} w(t_k,d_j) * Rel(t_k,c_i|d_j) \quad (6)$$

3.2 Multi-layer text classification framework

Multi-layer text classification framework is designed to handle large scale data with complex and high dimensions in a way of layer-by-layer. The proposed framework consists of two layers with three classifiers for three types of feature spaces. The first two SVM-NN classifiers in lower layer are applied on syntactic level and semantic level independently. Each document will be represented with two compacted vectors according to the similarity between document and all class centers in each classifier. These two compacted vectors are then combined to be the input of the third SVM-NN classifier which will output the final results. In the training stage of the efficient classification approach, the Support Vectors of each category are identified by using the conventional SVM training algorithm.In the nearest neighbor classification stage, the Euclidean distance formula is used to calculate the distance between the new input data point and the SVs from different categories. The category which has the shortest average distance between its SVs and the input data point is identified as the right category for the input data point.Fig.1. illustrates the Multi-layer text classification framework in detail. In the low layer, the first classifier is trained and tested using the documents which are represented by term-based representation.

3.2.1 Support vector machine training process

The support vector machine (SVM) has been reported as a discriminative classifier which is more accurate than most other classification models[8].The good characteristic of the SVM is due to the implementation of Structural Risk Minimization (SRM) principle, which entails finding an optimal separating hyper-plane as illustrated in Eq 7, thus providing the highly accurate classifier in most applications.

$$w.x+b=0 \quad (7)$$

The nearest data points to the optimal separating hyper-plane are called support vectors (SVs). The maximal margin can be found by minimizing ½ $||w||^2$.By minimizing, training data

points are separated and the optimal separating hyper-plane can be configured with the constraint[9] as illustrated in Eq 8.

$$Y_i(w.x_i + b) \geq 1 \quad (8)$$

3.2.2 Nearest neighbor classification approach

The nearest neighbor (NN) classification approach uses the nearest distance in determining the category of new vector. The unlabeled input data points are categorized to a particular category by finding the closet or distance from input data point and support vectors(SVs) of that particular category obtained from SVM training process. The KNN approach needs only a small number of training data points and this has contributed to the simplicity of the NN which makes it outperforms other classification approaches. The distance function for the NN classifier is the Euclidean distance formula and it is used to calculate the distance between the new unlabeled data point and the support vectors (SVs) obtained from SVM training process [10].

The next step is to calculate the distance between the new data point, P and one of the SVs, Q using the Euclidean distance formula. The Euclidean distance formula for this computation is illustrated in Eq. (9) where pi and qi are the coordinate of P or Q respectively, n is the total number of dimension for the data points, and D is the Euclidean distance between P and Q.

$$D = \sqrt{\sum_{i=1}^{n}(pi - qi)2} \quad (9)$$

The average distance of the SVs of a particular category and the new data point is calculated by using the formula as illustrated in Eq. (10), where N is the number of SVs for that particular category. The category which has the shortest average distance between its SVs and the new data point is identified as the right category for the new data point.

$$Davg = \frac{\sum_{i=1}^{n}(\sqrt{\sum_{i=1}^{n}(pi-qi)2})}{N} \quad (10)$$

According to the truth labels of training set and the predicted labels of test set of the first classifier, the center of each class can be determined by averaging the document vectors belonging to the class as showed in below Eq 11 as

$$Z_k = \sum_{dj \in ck} dj / |C_k| \quad (11)$$

Where $|C_k|$ is the number of documents in the kth class C_k. Based on the class centers, each document can be represented with a K dimension compressed vector $[s_{j1}, \ldots, s_{jK}]$ (K equals to the number of classes) where the value of the kth element is the similarity between document and the kth class center.

$$S_{jk} = d_j \cdot Z_k / \|d_j\| \|Z_k\| \quad (12)$$

Similarly, the second classifier is applied on the concept-based representation, to get the second K dimension compressed vector $[s_{j1}, s_{j2} \ldots s_{jk}]$ for each document.

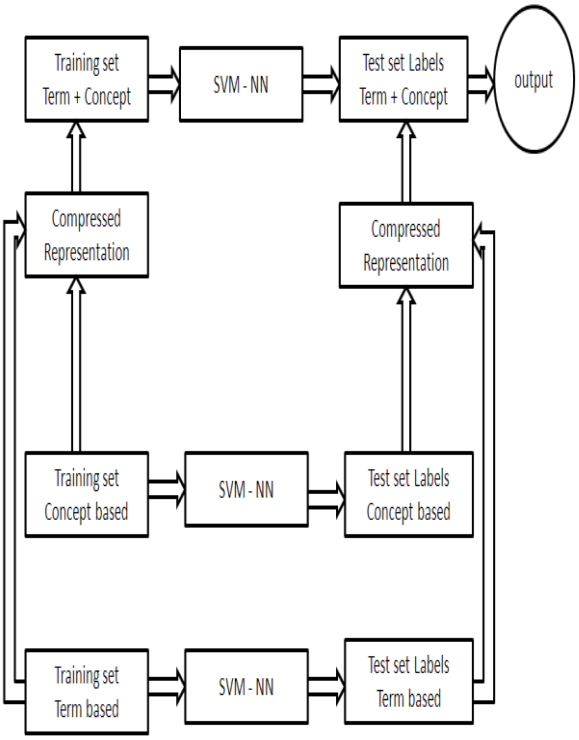

Fig 1: Multi- Layer Text Classification framework

Then, two K-dimension compressed vectors are combined as in equation 13 as

$$d_j = [s_{j1}, \ldots, s_{jk}, s_{j1}', \ldots, s_{jk}'] \quad (13)$$

s_{jk} is the similarity between the jth document represented in syntactic level and the kth class center obtained by the first classifier. s_{jk}' is the similarity between the jth document represented in semantic level and the kth class center obtained by the second classifier. This combined document representation will be the input of the third classifier. The primary information is effectively kept and the noise is reduced by compressing the original information, so that proposed model can guarantee the quality classification. Thus the final classification performance would be improved.

4. RESULTS

The proposed efficient text classification framework were tested on real data,20Newsgroups and Reuters-21578. Meanwhile, the experimental results also demonstrate that only using concepts or term to represent document usually gets worse performance than proposed model, because it cannot avoid inducing noises and losing information due to the limitation of background knowledge base and word sense disambiguation technique. The proposed model surpasses all the flat vector models and it is robust for classification algorithm. In addition, the results shown in Figs. 2 shows how far the proposed model increases text classification accuracy. Thus, our proposed method seems more useful in practice.

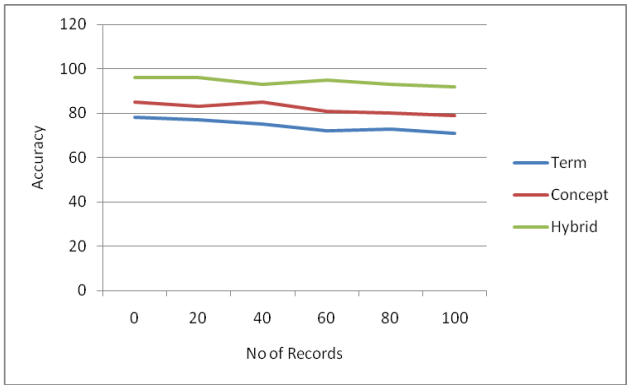

Fig. 2 Comparison of Performance

5. CONCLUSION

The proposed system represents document as two-level model with the aid of Wikipedia. In two-level representation model, levels are connected by the semantic relatedness between terms and concepts. A context-based method is adopted to identify the relatedness between terms and concepts by utilizing the link structure among Wikipedia articles. Multi-layer efficient text classification framework produces classification results with high accuracy. Experimental results on real data sets have shown that the proposed model and classification framework significantly improved the classification performance.

6. ACKNOWLEDGMENTS

Our thanks to staffs and students of Sri Ramakrishna Engineering College who gave their kind support towards development of the proposed system.

7. REFERENCES

[1] Hotho, A., Staab, S., and Stumme, G.2003.Wordnet improves textdocument clustering. In Proceedings of the semantic web workshop at the 26th ACM SIGIR ,pp 541–544.

[2] Gabrilovich, E., and Markovitch, S.2005. Feature generation for text categorization using word knowledge. In Proceedings of the 19 international joint conference on artificial intelligence, Edinburgh ,pp 1048–1053.

[3] Gabrilovich, E., and Markovitch, S.2006. Overcoming the brittleness bottleneck using Wikipedia Enhancing text categorization with encyclopedic knowledge.In Proceedings of the 21st AAAI, Boston, MA, USA ,pp 1606–1611.

[4] Banerjee, S., Ramanathan, K., and Gupta, A.2007.Clustering short texts using Wikipedia.InProceedings of the 30th ACM SIGIR ,pp 787–788.

[5] Hu, J., Fang, L., Cao, Y., Zeng, H., Li,H.,Yang, Q., and Chen,Z.2008.Enhancing text clustering by leveraging wikipedia semantics. In Proceedings of the 31st ACM SIGIR ,pp 179 -186.

[6] Wang, P., and Domeniconi, C.2008.Building semantic kernels for text classification using Wikipedia. In Proceedings of the 14th ACM SIGKDD, New York, NY, USA,pp 713–721.

[7] Chang, M., and Roth, D.2008. Importance of semantic representation: Dataless classification.In Proceedings of th 23rd AAAI conference on artificial intelligence , pp 830–835.

[8] Isa, D., Lee, L. H., Kallimani, V. P., and Rajkumar, R. 2008.Text document preprocessing with the Bayes formula for classification using the support vector machine. IEEE Transactions on Knowledge and Data Engineering, 20(9),pp1264–1272.

[9] Joachims, T.2008. Text categorization with support vector machines: learning with many relevant features. In Proceedings of the 10th European conference on machine learning (ECML 98), pp 137–142.

[10] Han, E. H., Karypis, G. & Kumar, V.2008.Text categorization using weighted adjusted K-nearest neighbor classification.Technical Report, Department of Computer Science and Engineering, Army HPC Research Centre, University of Minnesota, Minneapolis, USA.