IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 4, Ver. VI (Jul – Aug. 2014), PP 14-26
www.iosrjournals.org

A novel semantic level text classification by combining NLP and Thesaurus concepts

R. Nagaraj[1], Dr.V. Thiagarasu[2], P.Vijayakumar[3]
Research Scholar, Karpagam University, Coimbatore, India
Associate Professor of Computer Science, Gobi Arts and Science College, Gobichettipalayam, India
MPhil Scholar, Kaamadhenu Arts and Science College, Sathyamangalam, India

Abstract: *Text categorization (also known as text classification or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, but it can be too expensive or simply not feasible given the time constraints of the application or the number of documents involved. In the previous approaches only the Wikipedia concepts related to terms in syntactic level are used to represent document in semantic level. This paper proposes a new approach to represent semantic level with the use of Word Net. The semantic weight of terms related to the concepts from Wikipedia and Word Net are used to represent semantic information. The semantic vector space model of terms by combining the Word Net and Wikipedia is being further improved the classification accuracy of the Text classification. Because of, two different concept extractor are gives the concepts related to the terms in the syntactic level o find the better concept vector space for documents. So we obtain the improved classification by using this approach. In this study the classification framework are presented. In classification framework, the primary information is effectively kept and the noise is reduced by compressing the original information, so that this framework can guarantee the quality of the input of all classifiers. This proposed method can help to further improve the performance of classification framework by introducing Wikipedia with Word Net. We find that the proposed approach result in a high classification accuracy.*

Keywords**:** *Text classification, vector space model, Wikipedia, Word Net.*

I. Introduction

Data mining is the way to help organization make full use of the data stored in their databases and when it comes to decision making, this is true in all fields and all different types of organizations. Data mining is the task of discovering interesting and hidden patterns from large amounts of data where the data can be stored in databases, data warehouses, OLAP (online analytical process) or other repository information. It is also defined as knowledge discovery in databases (KDD).

Text mining, roughly equivalent to text analytics and it refers to the process of deriving high-quality information from text. Types of text mining tasks include text clustering, text categorization, concept/entity extraction, sentiment analysis, production of granular taxonomies, document summarization, and entity relation modeling. Text analysis involves information retrieval, to study word frequency distributions lexical analysis, pattern recognition, information extraction, tagging/annotation, data mining techniques including link and association analysis, predictive analytics and visualization. The main goal is, to turn text into data for analysis, through application of natural language processing (NLP) and analytical methods. Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. An organization can successfully use text analytics to gain insight into content-specific values such as intensity, sentiment, emotion, and relevance with an iterative approach. Reason for that text analytics technology is still considered to be an emerging technology.

Apart from manual classification and hand-crafted rules, there is a third approach to text classification called machine learning-based text classification. In ML (machine learning), the set of rules or, more commonly, the decision criterion of the text classifier, is acquired automatically from training data. This technique is also called statistical text classification if the learning method is statistical. In statistical text classification, we need a number of good example documents (or training documents) for each class. The necessary for manual classification is not eliminated because the training documents come from a person who has labeled them - where labeling refers to the process of annotating each document. But labeling is arguably an easier task than writing rules. Mostly anybody can look at a document and decide whether or not it is related to China. Formerly such labeling is already implicitly part of an existing workflow. For instance, you may go through the news articles returned by a standing query each morning and give relevance feedback by moving the relevant articles to a special folder like multicore-processors.

II. Related Works

VSM (Vector Space Model) is the most popular document representation model for text clustering, classification and information retrieval. In early literature, term-based VSM, representing one document as a term vector, was widely used. The weight of each term in a document is usually measured via two schemes: Binary (1 for term appearing in the document, 0 for not) and Term Frequency-Inverse Document Frequency (TF-IDF). However, both approaches only contain the literal information in document. Some methods were proposed to mine the underlying semantic structure in textual data, such as Latent Dirichlet Allocation (LDA) in Blei, Ng, & Jordan[1] and Latent Semantic Indexing (LSI) in Deerwester, Dumais, Landauer, Furnas, & Harshman[2]; Hotho et al. [3] took 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. Thus, many researches began to make use of Wikipedia, the largest electronic encyclopedia to date. Nouali & Blache [4]. To some extent, these methods make up for the shortage of term-based VSM, but they cannot discover as much semantic information as described in text data only by analyzing syntactic information via statistic methods. Syed, Finin, and Joshi[5] was interested in finding semantically related concepts which were also common to a set of documents.

In Wang, Hu, Zeng, Chen, and Chen [6] 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. 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 relatedness to change the concept weight. In the existing paper, concept weight is effected by the semantic relatedness between concept and the given document, which is equal to the average relatedness between concept and other concepts (contextual concepts) within the document. Here, the semantic relatedness measure between concepts also adopted link-based concept relatedness method Milne and Witten [7], Medelyan, Witten, and Milne [8]. In Huang et al. [9] compared three models (concept-based VSM, Term + Concept VSM and Replaced VSM) with term-based VSM. In the experiments, they used the WordNet and Wikipedia as the background knowledge bases respectively. Experimental results showed that Term + Concept VSM usually can improve successfully the performance in text clustering and concept-based VSM did not perform better than term-based VSM in most cases. These observations gave us a hint: concept-based VSM can supply more information for discriminating documents, but only using concepts cannot represent document sufficiently. Concept mapping could result in loss of information or addition of noise. It is necessary to include both term and concept in representation model. In order to make use of term and concept information in text classification and clustering tasks, an alternative method is to liner combining the similarity values which are calculated based on term-based VSM and concept-based VSM respectively. However, as shown in the literatures, this method depends on the input parameters.

Huang, Milne, Frank, & Witten [10] mapped candidate phrases in the given document to Wikipedia articles by leveraging an informative and compact vocabulary – the collection of anchor texts in Wikipedia. The existing adopted method is more similar with Huang et al.[10] used where Wikipedia's anchor text vocabulary is used to connect terms to Wikipedia articles. In this way the number of concepts in a document is no more than the number of terms. Meanwhile, different terms with the same meaning might be mapped to the same Wikipedia article because anchors linked to the same article are also often couched in different words. In Jing, Zhou, Ng, and Huang [11] implicitly embedded the semantic information to document representation via kernel method by multiplying document-term tf-idf matrix and term similarity matrix, where the term similarity was computed based on Word Net.

In Hu et al.[12] built document-concept matrix through exact-match and relatedness-match which requires to compute the tf-idf value of term in the whole Wikipedia article collection. In Gabrilovich and Markovitch [13], [14],[15] 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. Besides identifying the related concepts, weighting the concepts is also a vital technology to build concept-based VSM. It is time consuming. Banerjee, Ramanathan, and Gupta [16] treated the entire document as query strings to Wikipedia and associate the document with the top articles in the returned result list. Due to the limited background knowledge and concept mapping technology, extracted concepts might not contain the term information exactly and completely. Many Researchers began to use both term and concept information to represent document, for instance, Term + Concept VSM and Replaced VSM. The Replaced VSM represents document with concepts and terms which do not have any related concept in knowledge base of Wang et al.,[17].

2.1 Introduction to Word Net

The lexical database Word Net is particularly well suited for similarity measures, because it organizes nouns and verbs into hierarchies of is–a relations. In version 2.0, there are nine noun hierarchies that include 80,000 concepts, and 554 verb hierarchies that are made up of 13,500 concepts. Is–a relations in WordNet do not cross part of speech boundaries, so Word Net–based similarity measures are limited to making judgments between noun pairs (e.g., cat and dog) and verb pairs (e.g., run and walk). While WordNet includes adjectives and adverbs, these are not organized into is–a hierarchies so similarity measures cannot be applied. However, concepts can be related in many ways beyond being similar to each other. For example, a wheel is a part of a car, night is the conflicting to day, snow is made up of water, a knife is used to cut bread, and so forth. As such Word- Net provides additional (non–hierarchical) relations such as has–part, is–made–of, is–an–attribute–of, etc. In addition, each concept (or word sense) is described by a short written definition or gloss. Measures of relatedness are based on these additional sources of information, and as such can be applied to a wider range of concept pairs. For example, they can cross part of speech boundaries and assess the degree to which the verb murder and the noun gun are related. They can even measure the relatedness of concepts that do not reside in any is–a hiearchy, such as the adjectives violent and harmful.

As Pucher [18] has shown different Word Net- based measures and contexts are best for word prediction in conversational speech. The JCN measure performs best for nouns using the noun-context. The LESK measure performs best for verbs and adjectives using a mixed word-context. In Demetriou et al.,[19] generated N-best lists from phoneme confusion data acquired from a speech recognizer, and a pronunciation lexicon. Then sentence hypotheses of varying Word-Error-Rate (WER) were generated based on sentences from different genres from the British National Corpus (BNC). It was shown by them that the semantic model can improve recognition, where the amount of improvement varies with context length and sentence length. Thereby it was shown that these models can make use of long-term information. Most of the work dealing with relatedness and similarity measures has been developed using WordNet. While WordNet represents a well structured taxonomy organized in a meaningful way, questions arise about the need for a larger coverage. E.g., WordNet 2.1 does not include information about named entities such as Condoleezza Rice, Salvador Allende or The Rolling Stones as well as specialized concepts such as exocytosis or P450.

2.2 Introduction to Wikipedia

In contrast, Wikipedia provides entries on a vast number of named entities and very specialized concepts. The English version, as of 14 February 2006, contains 971,518 articles with 18.4 million internal hyperlinks, thus providing a large coverage knowledge resource developed by a large community, which is very attractive for information extraction applications [20]. Also, it provides also taxonomy by means of its categories: articles can be assigned one or more categories, which are further categorized to provide a category tree. In practice, the taxonomy is not designed as a strict hierarchy or tree of categories, but allows multiple categorization schemes to co-exist simultaneously. As of January 2006, 94% of the articles have been categorized into 91,502 categories. The strength of Wikipedia lies in its size, which could be used to overcome current knowledge bases' limited coverage and scalability issues. Such size represents on the other hand a challenge: the search space in the Wikipedia category graph is very large in terms of depth, branching factor and multiple inheritance relations, which creates problems related to finding efficient mining methods.

In addition, the category relations in Wikipedia cannot only be interpreted as corresponding to is-a links in taxonomy since they denote meronymic relations as well. As an example, the Wikipedia page for the Nigerian musician Fela Kuti belongs not only to the categories MUSICAL ACTIVISTS and SAXOPHONISTS (is-a) but also to the 1938 BIRTHS (has-property) [21]. This is due to the fact that, rather than being a well-structured taxonomy, the Wikipedia category tree is an example of a folksonomy, namely a collaborative tagging system that enables the users to categorize the content of the encyclopedic entries. Folksonomies as such do not strive for correct conceptualization in contrast to systematically engineered ontologies. They rather achieve it by collaborative approximation.

III. Semantic Level Text Classification By Thesaurus Concepts

Two-level Representation Model (2RM) that represents syntactic information and semantic information with two levels. Term-based VSM 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. The key technique to build 2RM model is to construct the semantic level 2RM represents document in a two-level vector space containing syntactic (term) and semantic (related concept) information respectively.

3.1 Two-level representation model

In this section, Two-level Representation Model (2RM) that represents syntactic information and semantic information with two levels. Term-based VSM 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. The key technique to build 2RM model is to construct the semantic level. In this paper, a context-based method is proposed to find the most relevant concept for each term based on the document structure information (e.g., document-paragraph) and Wikipedia link structure.

The semantic relatedness between term and its candidate concepts in a given document is computed according to the context information as follows (1).

$$\text{Rel}(t, c_i|d_j) = \frac{1}{|T|-1} \sum \frac{1}{|cs_l|} \sum \text{SIM}(c_i, c_k) \quad (1)$$

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

$$\text{SIM}(c_i,c_k) = 1 - \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \quad (2)$$

where A and B are the sets of all articles that link to concepts c_i and c_k respectively, and W is the set of all articles in Wikipedia. The equation (2) is based on term occurrences on Wikipedia-pages. Pages that contain both terms indicate relatedness, while pages with only one of the terms suggest the opposite.

Higher value of $\text{Rel}(t, c_i|d_j)$ means that concept c_i is more semantically related to term t, because c_i is much more similar to the relevant concepts of other terms in d_j (such terms are the context of term t). The concepts with highest relatedness will be used to properly build the concept vector in semantic level, i.e., each term will be finally mapped into its most related concept. Based on $\text{Rel}(t, c_i|d_j)$ and term's weight $w(t_k, d_j)$, the concept's weight is defined as their weighted sum as follows(3).

$$W(c_i , d_j) = \sum w(t_k, d_j) * \text{Rel}(t_k, c_i|d_j) \quad (3)$$

Different terms may be mapped to a same concept, and some term such as ''dealt'' has no concept in Wikipedia. Because of these many-to-one mapping, the synonym information can be considered in our proposed 2RM model. In order to deal with the second situation, some terms do not have related concept, a multi-layer classification framework is designed, to make use of term and concept information during the classification processing.

3.2 Multi-layer classification framework

In this step presents constructing the MLCLA framework. MLCLA framework includes two classification procedures in low layer; they can be implemented in series or parallel. When running in series; two data matrices based on different representation levels syntactic and semantic) can be loaded one by one. Therefore, the required memory space depends on the larger matrix plus compressed representation matrix, rather than the summation of term-based matrix and concept-based matrix. On the other hand, when running in parallel, two classifiers in low layer can be built at the same time, and the classifier in high layer is very fast on the basis of low dimension compression space. Now further analyze the time complexity of MLCLA. N represents the number of documents, M denotes the number of terms or concepts in document collection, K is the number of classes and m is the average number of terms or concepts in one document. The low layer of MLCLA includes two classification procedures, based on syntactic level and semantic level respectively.

In the low layer, the first classifier is trained and tested using the documents which are represented by term-based VSM, i.e., the syntactic information in 2RM model. 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 this class.

$$Z_k = \frac{\sum dj}{|ck|} \quad (4)$$

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 $[S_{j1},\ldots\ldots,S_{jK}]$ (K equals to the number of classes) where the value of the kth element is the similarity between document and the kth class center.

$$S_{jK} = \frac{dj \bullet Zk}{||dj|| \, ||Zk||} \quad (5)$$

Similarly, the second classifier is applied on the concept-based VSM, i.e., the semantic information in 2RM model, to get the second K dimension compressed vector $[S'_{j1},\ldots\ldots,S'_{jK}]$ for each document. Then, two K-dimension compressed vectors are combined as follows (6).

$$d_j = [S_{j1},\ldots\ldots,S_{jK},S'_{j1},\ldots\ldots,S'_{jK}] \quad (6)$$

S_{jK} is the similarity between the jth document represented in syntactic level of the 2RMmodel and the kth class center obtained by the first classifier. S'_{jK} is the similarity between the jth document represented in

semantic level of the 2RM model and the kth class center obtained by the second classifier. This combined document representation will be the input of the third classifier in the high layer of MLCLA.

In MLCLA framework, the primary information is effectively kept and the noise is reduced by compressing the original information, so that MLCLA can guarantee the quality of the input of all classifiers. Thus we believe the final classification performance would be improved. Because MLCLA framework includes two classification procedures in low layer, they can be implemented in series or parallel. When running in series, two data matrices based on different representation levels (syntactic and semantic) can be loaded one by one. Therefore, the required memory space depends on the larger matrix plus compressed representation matrix, rather than the summation of term-based matrix and concept-based matrix. On the other hand, when running in parallel, two classifiers in low layer can be built at the same time, and the classifier in high layer is very fast on the basis of low dimension compression space.

IV. A Novel Semantic Level Text Classification By Combining Nlp And Thesaurus Concepts

In this section introduces a measure of relateness besed on formulation of information content, which is a value that is assigned to each concept in a hierarchy based on evidence found in a corpus. Before describing this measure of relatedness we first introduce the notion of information content, which is simply a measure of the specificity of a concept. A concept with a high information content is very specific to a particular topic, when concepts with lower information content are associated with more general and less specific concepts. Thus, carving fork has a high information content while entity has low information content.

Information content of a concept is estimated by counting the frequency of that concept in a large corpus and thereby determining its probability via a maximum likelihood estimate. According to this, the negative log of this probability determines the information content of the concept(7):

$$IC(concept) = - \log(P(concept)) \quad (7)$$

If sense-tagged text is available, we can be attained the frequency counts of concepts directly, since each concept will be associated with a unique sense. If sense-tagged text is not available it will be necessary to adopt an alternative counting scheme. In this technique counting the number of occurrences of a word type in a corpus, and then by using the number of different concepts/senses associated with that word, dividing that count. This value is then assigned to each concept.

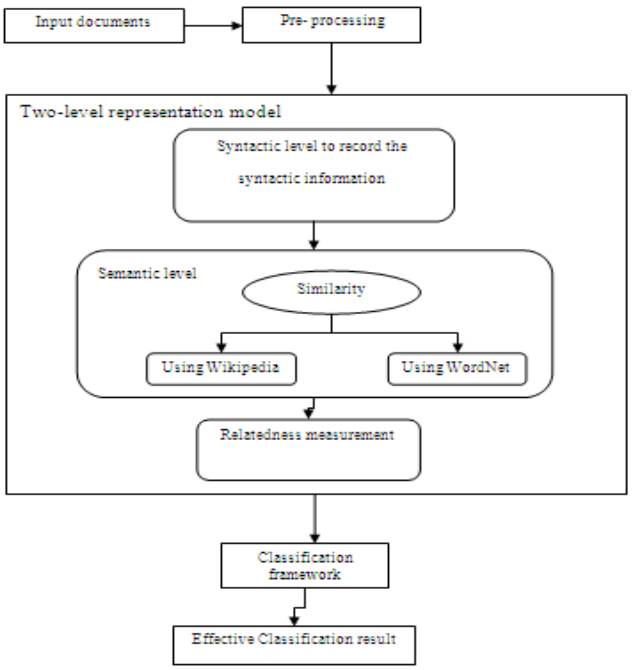

Fig 1. Overall architecture diagram

For example, suppose that the word type bank occurs 20 times in a corpus and there are two concepts associated with this type in the hierarchy, one for river bank and the other for financial bank. Each of these concepts would receive a count of 10. If the occurrences of bank were sense tagged then the relevant counts could simply be assigned to the appropriate concept. In this method we choose to assign the total count to all the concepts and not divide by the number of possible concepts. Thus we would assign 20 to river bank and finanvial bank in the example above. This decision was based on the observation that by distributing the

frequency count over all the concepts associated with a word type we effectively assign a higher relative frequency to those words having fewer senses. This would lead us to estimate higher probability and therefore assign a lower value of information content to such concepts.

Regardless of how they are counted, the frequency of a concept includes the frequency of all its subordinate concepts since the count we add to a concept is added to its subsuming concept as well. Note that the counts of more specific concepts are added to the more general concepts, which is not from the more general to specific. Thus, counts of more specific concepts percolate up to the top of the hierarchy, incrementing the counts of the more general concepts as they proceed upward. As a result, concepts that are higher up in the hierarchy will have hihger counts than those at lower more specific levels and have higher probabilities associated with them. Such high probability concepts will have low values of information content since they are associated with more general concepts.

This measure of semantic similarity uses the information content of concepts along with their positions in the noun is – a hierarchies of Word Net to compute a value for the semantic relatedness of the concepts. The principle idea behind this measure of semantic relatedness is that two concepts are semantically relatd proportional to the amount of information they share in common. The quantity of information common to two concepts is determined by the information content of the lowest concepts in the hierarchy that subsumes both the concepts. This concepts is known as the lowest common subsumer of the two concepts. Thus, the measure of similarity is defined as folows(8):

$$SIMres(c_1,c_2) = IC(lcs(c_1,c_2)) \quad (8)$$

We note that this measure does not consider the information content of the concept themselves, nor does it directly consider the path length.

$$SIM^*(c_j,c_k) = \max_{C \in S(cj,ck)} [- \log(P(c)] \quad (9)$$

Where $S(c_j,c_k)$ is the set of concepts that subsume both c_j and c_k. Notice that although similarity is computed by considering all upper bounds for the two concept but the information measure has the effect of identifying minimal upper bounds, because no class is less informative than its superordinates.

The semantic relatedness between term and its candidate concepts in a given document is computed according to the context information as follows(10).

$$Rel(t, c_i|d_j) = \frac{1}{|T|-1} \sum \frac{1}{|csl|} \sum SIM(c_i,c_k)\ SIM^*(c_j,c_k) \quad (10)$$

where T is the term set of the jth document d_j , t_l is a term in d_j except for t and 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 and $SIM^*(c_j,c_k)$ is the semantic relatedness between two concepts, which is calculated with the WordNet hyperlinks.

$$SIM(c_i,c_k) = 1 - \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \quad (11)$$

where A and B are the sets of all articles that link to concepts ci and ck respectively, and W is the set of all articles in Wikipedia. The equation (11) is based on term occurrences on Wikipedia-pages. Pages that contain both terms indicate relatedness, while pages with only one of the terms suggest the opposite.

Higher value of $Rel(t, c_i|d_j)$ means that concept c_i is more semantically related to term t, because c_i is much more similar to the relevant concepts of other terms in d_j (such terms are the context of term t). The concepts with highest relatedness will be used to properly build the concept vector in semantic level, i.e., each term will be finally mapped into its most related concept. 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(12).

$$W(c_i , d_j) = \sum w(t_k, d_j) * Rel(t_k, c_i|d_j) \quad (12)$$

Different terms may be mapped to a same concept, and some term such as ''dealt'' has no concept in Wikipedia. Because of these many-to-one mapping, the synonym information can be considered in our proposed 2RM model. In order to deal with the second situation, some terms do not have related concept, a multi-layer classification framework is designed, to make use of term and concept information during the classification processing.

4.1 Classification framework

In this step we are constructing the classification framework. This classification framework includes two classification procedures in low layer; they can be implemented in series or parallel. When running in series; two data matrices based on different representation levels syntactic and semantic) can be loaded one by one. Therefore, the required memory space depends on the larger matrix plus compressed representation matrix, rather than the summation of term-based matrix and concept-based matrix. On the other hand, when running in parallel, two classifiers in low layer can be built at the same time, and the classifier in high layer is very fast on the basis of low dimension compression space. Now we further analyze the time complexity of classification framework. N represents the number of documents, M denotes the number of terms or concepts in document collection, K is the number of classes and m is the average number of terms or concepts in one document. The

low layer of this framework includes two classification procedures, based on syntactic level and semantic level respectively.

In the low layer, the first classifier is trained and tested using the documents which are represented by term-based VSM, i.e., the syntactic information in 2RM model. 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 this class.

$$Z_k = \frac{\sum dj}{|ck|} \quad (13)$$

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 $[S_{j1},\ldots\ldots,S_{jK}]$ (K equals to the number of classes) where the value of the kth element is the similarity between document and the kth class center.

$$S_{jK} = \frac{dj \bullet Zk}{||dj|| \, ||Zk||} \quad (14)$$

Similarly, the second classifier is applied on the concept-based VSM, i.e., the semantic information in 2RM model, to get the second K dimension compressed vector $[S'_{j1},\ldots\ldots,S'_{jK}]$ for each document. Then, two K-dimension compressed vectors are combined as follows (15).

$$d_j = [S_{j1},\ldots\ldots,S_{jK},S'_{j1},\ldots\ldots,S'_{jK}] \quad (15)$$

S_{jK} is the similarity between the jth document represented in syntactic level of the 2RM model and the kth class center obtained by the first classifier. S'_{jK} is the similarity between the jth document represented in semantic level of the 2RM model and the kth class center obtained by the second classifier. This combined document representation will be the input of the third classifier in the high layer of classification framework.

In classification framework, the primary information is effectively kept and the noise is reduced by compressing the original information, so that this framework can guarantee the quality of the input of all classifiers. Thus we believe the final classification performance would be improved. Because classification framework includes two classification procedures in low layer, they can be implemented in series or parallel. When running in series, two data matrices based on different representation levels (syntactic and semantic) can be loaded one by one. Therefore, the required memory space depends on the larger matrix plus compressed representation matrix, rather than the summation of term-based matrix and concept-based matrix. On the other hand, when running in parallel, two classifiers in low layer can be built at the same time, and the classifier in high layer is very fast on the basis of low dimension compression space.

V. Performance Evaluation

5.1 Data set

The proposed representation model and classification framework were tested on three real data, 20Newsgroups, Reuters-21578 and Classic3. Six subsets were extracted from 20Newsgroups: 20NGDiff4, 20NG-Sim4, 20NG-Binary, 20NG-Multi5, 20NG-Multi10 and 20NG-Long.

Table 1: 20NewsGroup subsets

Dataset	Categories
20NG-Binary	talk.politics.mideast, talk.politics.misc
20NG-Multi5	comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space, talk.politics.mideast
20NG-Multi10	alt.atheism, comp.sys.mac.hardware, misc.forsale, rec.autos, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, talk.politics.guns
20NG-Diff4	comp.graphics, rec.sport.bassball, sci.space, talk.politics.mideast
20NG-Sim4	comp.graphics, comp.os.ms-windows.misc, rec.autos, sci.electronics
20NG-Long	comp/, sci/, talk/

Tables 1 and 2 list the categories and the number of documents contained in these subsets. In this paper, 20NG-Long is a collection of long documents containing three categories ‘‘comp’’, ‘‘sci’’ and ‘‘talk’’. In each category, 70 documents with the most large size were extracted from the corresponding topic in 20Newsgroups (documents from topic ‘‘rec’’ were not included because there are few long documents in ‘‘rec/’’). In 20NG-long, the minimal document’s size is 10 K, the maximal one is 158 KB and the average size is 29 KB. Another two data subsets were created from Reuters-21578: R-Min20Max200 and R-Top10. R-Min20Max200 consists of 25 categories with at least 20 and at most 200 documents, 1413 documents totally. In R-Top10, 10 largest categories were extracted from the original data set including 8023 documents. For Classic3, the whole dataset was used in the experiment.

Table 2: Data set summary

Dataset	Classes	Documents	Words	Concepts
20NG-Binary	2	500	3376	2987
20NG-Multi5	5	500	3310	2735
20NG-Multi10	10	500	3344	2772
20NG-Diff4	4	4000	5433	4362
20NG-Sim4	4	4000	4352	3502
20NG-Long	3	210	4244	3738
R-Min20Max200	25	1413	2904	2450
R-Top 10	10	8023	5146	4109
Classic3	3	3891	4745	3737

For our experiment we consider the two subset from 20Newsgroups dataset, one subset created from Reuters-21578 and Classic3. We are taking 20NG-Multi10 subset and 20NG-Sim4 subset which is extracted from the 20Newsgroups. 20NG-Multi10 dataset consists of totally 500 documents and 10 classes. In 20NG-Multi10, 3344 words and 2772 concepts are there. As well as 20NG-Sim4 consists of 4000 documents and 4 classes. 4352 words and 3502 concepts are there in this dataset. From the Reuters-21578, we are taking the R-Top 10 subset which consists of 10 classes and 8023 documents. In this dataset, 4109 concepts and 5146 words are there. For the Classic 3, 3 classes and 3891 documents are present. This dataset consists of 4745 words and 3737 concepts. In this paper, we only consider the single-label documents. Wikipedia and WordNet are used as background knowledge. Wikipedia contains 2,388,612 articles (i.e., concepts) and 8,339,823 anchors in English and WordNet in version 2.0, there are nine noun hierarchies that include 80,000 concepts, and 554 verb hierarchies that are made by 13,500 concepts.

Table 3: F-measure result from the experiment of the dataset

Dataset	Training set size of dataset					
	1/2	1/4	1/6	1/8	1/10	1/12
20NG-Multi10	0.92	0899	0.903	0.895	0.883	0.88
20NG-Sim4	0.91	0.88	0.873	0.861	0.85	0.85
R-Top10	0.92	0.85	0.79	0.79	0.69	0.73
Classic3	0.98	0.93	0.88	0.86	0.8	0.8

From Table 2 we can see the number of words and concepts extracted from each data set. The words were extracted by preprocessing steps, selecting only alphabetical sequences, stemming them, removing stop words and filtering them by the document frequency. Then, we determined the Wikipedia and WordNet concepts for these words in each document via the method (Note: once a word was stemmed, its original form was used to correctly identify relevant Wikipedia and WordNet concept). Table 2 shows that the number of distinct concepts appearing in a data set is usually lower than the number of words. Meanwhile, parts of words (about 10 percent) do not have relevant concepts. The main one is to test the performance of proposed 2RM model and classification framework on real datasets by comparing with various flat document representation models plus basic classification algorithm (e.g., SVM or KNN). Table 3 shows the F-measure result from the experiment of the dataset. It shows the F-measure is improved in the proposed system compared to the existing system.

5.2 Accuracy comparison

In this section performance is evaluated in terms of accuracy. In this graph we have taken the parameters called accuracy and training set size of four dataset namely 20NG-Multi10, 20NG-Sim4, R-Top10 and Classic3. It helps to analyze the existing system and proposed combining technique. Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

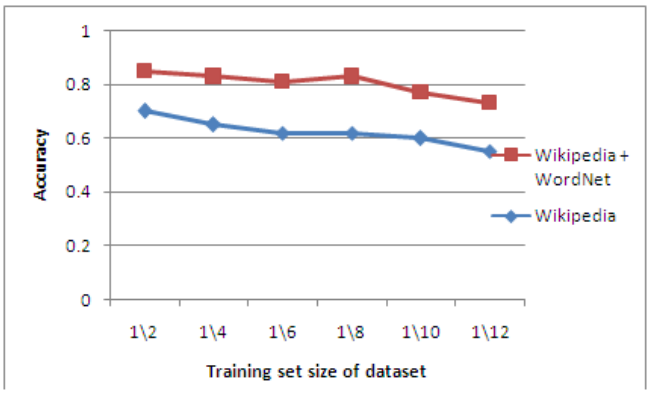

Fig 2. Accuracy comparison for 20NG-Multi10

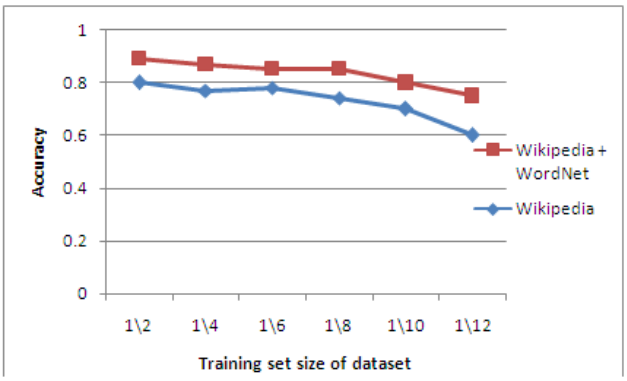

Fig 3. Accuracy comparison for 20NG-Sim4

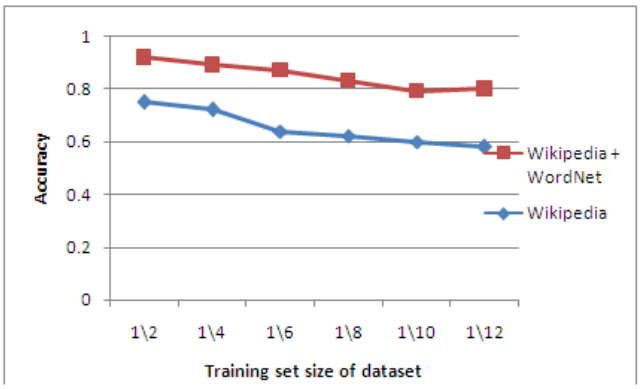

Fig 4. Accuracy comparison for R-Top10

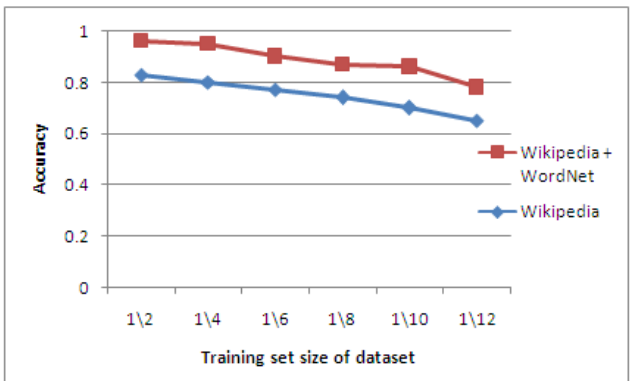

Fig 5. Accuracy comparison for Classic3

The Accuracy parameter will be the Y axis and training set size of dataset will be the X axis. Then we compare the accuracy performance. From this graph we identify that the accuracy of the proposed system is higher than the existing system. From this we easily understood the proposed system has more effective than exiting one.

5.3 Precision comparison

In this section performance is evaluated in terms of precision. Graph gives the precision comparison between the existing and proposed. It can be defined as

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

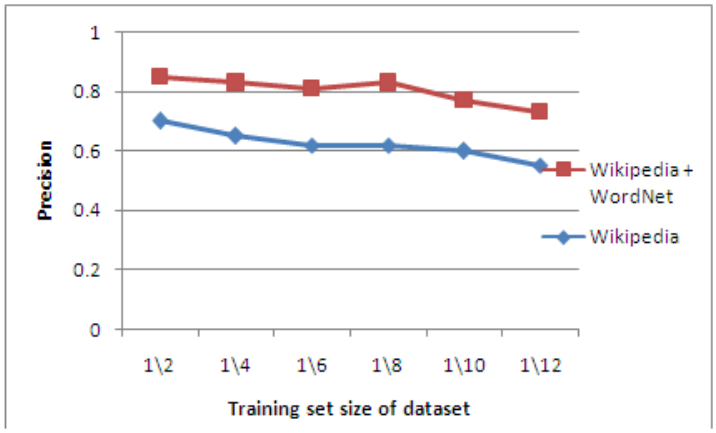

Fig 6. Precision comparison for 20NG-Multi10

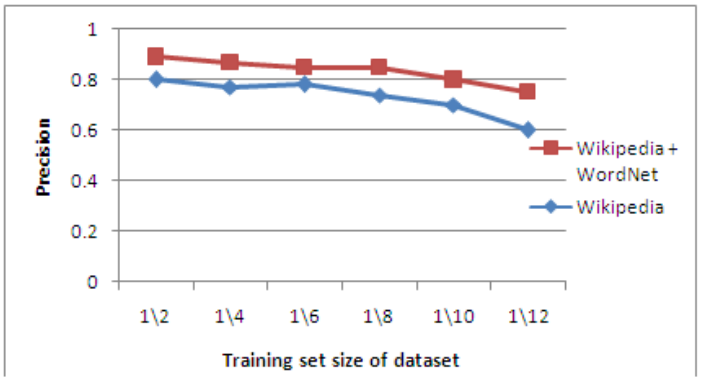

Fig 7. Precision comparison for 20NG-Sim4

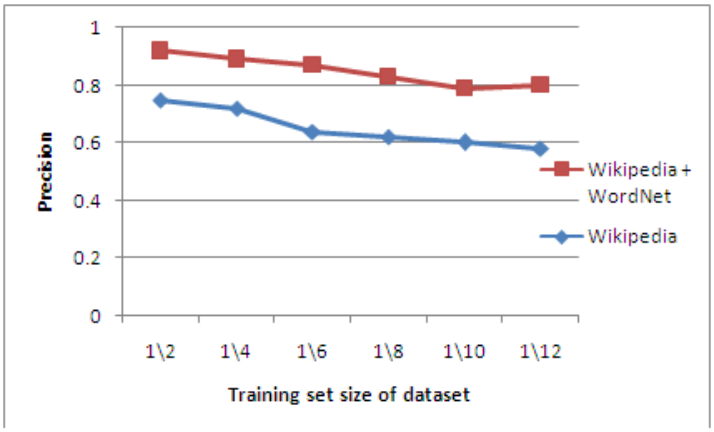

Fig 8. Precision comparison for R-Top10

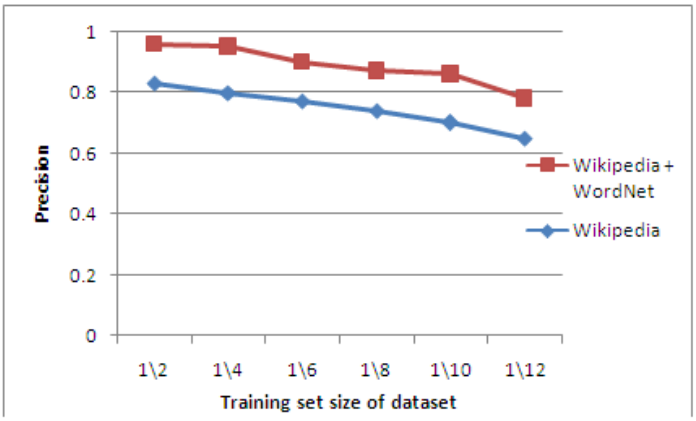

Fig 9. Precision comparison for Classic3

In the graph X-axis will be training set size of dataset of four dataset namely 20NG-Multi10, 20NG-Sim4, R-Top10 and Classic3 and Y-axis will be precision parameter. In the data sets also our proposed system has more precision compare to existing system. From this graph, proposed paper has effective in precision parameter.

5.4 F-measure comparison

F-measure distinguishes the correct classification of document labels within different classes. In essence, it assesses the effectiveness of the algorithm on a single class, and the higher it is, the better is the clustering. It is defined as follows:

$$F=2.precision.recall/precision+recall$$

then,

$$F(i,j)=2PR/P+R \Rightarrow F_c=\sum i(|i|*F(i))/\sum i_{|i|}$$

where for every class i is associated a cluster j which has the highest F-measure, F_c represents the overall F-measure that is the weighted average of the F-measure for each class i and |i| is the size of the class.

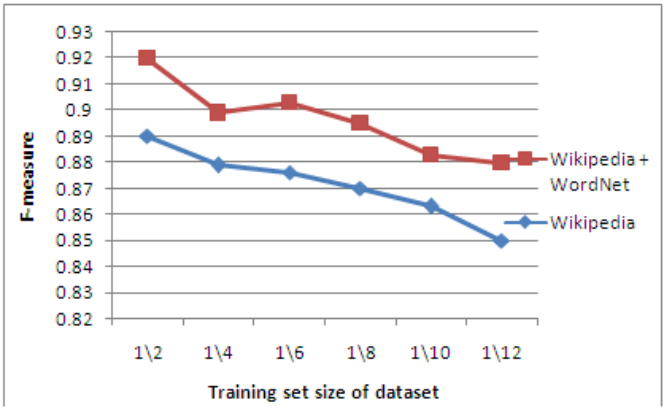

Fig 10. F-measure comparison for 20NG-Multi10

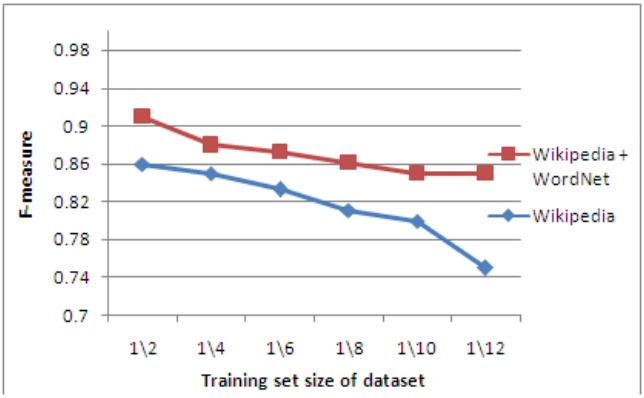

Fig 11. F-measure comparison for 20NG-Sim4

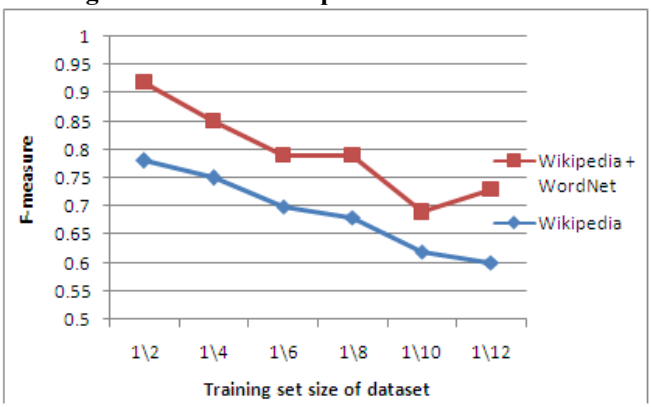

Fig 12. F-measure comparison for R-Top10

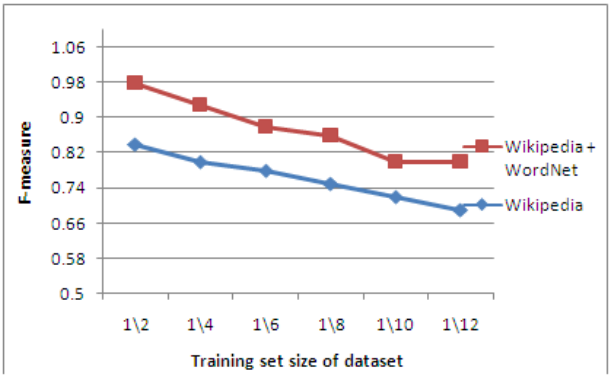

Fig 13. F-measure comparison for Classic3

The F-measure parameter will be the Y axis and training set size of dataset of four dataset namely 20NG-Multi10, 20NG-Sim4, R-Top10 and Classic3 will be the X axis. Then we compare the F-measure performance. From this graph we identify that the F-measure of the proposed system is higher than the existing system. From this we easily understood the proposed system has more effective than exiting one.

VI. Conclusion

In this work, we represent document as a two-level model with the aid of WordNet and Wikipedia. In the two-level representation model, one for term information, the other for concept information and these 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 WordNet and Wikipedia articles, which is also used to select the most appropriate concept for a term in a given document. By combining the WordNet and Wikipedia, we get the improved performance of the existing paper. Based on the two-level representation model, we propose a classification framework to analyze text data. By introducing the combining technique of WordNet and Wikipedia accuracy rate will be increased and error rate is decreased observed from the experiment.

References

[1]. D. Blei, A. Ng, and M. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research, 3, (2003), 993–1022.
[2]. S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, Indexing by latent semantic analysis, Journal of the Society for Information Science, 41(6), (1990), 391–407.
[3]. A. Hotho, S. Staab, and G. Stumme, Wordnet improves text document clustering. In Proceedings of the semantic web workshop at the 26th ACM SIGIR, (2003), 541–544.
[4]. O. Nouali, and P. Blache, A semantic vector space and features-based approach for automatic information filtering, Expert Systems with Applications, 26(2), (2004), 171–179.
[5]. Z. Syed, T. Finin, andA. Joshi, Wikipedia as an ontology for describing documents, In Proceedings of the 2nd international conference on weblogs and social media, Washington, (2008), 136–144.
[6]. P. Wang, J. Hu, J. Zeng, L. Chen, and Z. Chen, Improving text classification by using encyclopedia knowledge. In Proceedings of the 7th ICDM. Omaha, NE, USA, (2007),pp. 332–341.
[7]. D. Milne, and I. Witten, An effective, low-cost measure of semantic relatedness obtained from wikipedia links, In Proceedings of the workshop on Wikipedia and artificial intelligence at AAAI,(2008), 25–30.
[8]. O. Medelyan, I. Witten, and D. Milne, Topic indexing with wikipedia, In Proceedings of the AAAI wikipedia and AI workshop, (2008).

[9]. A. Huang, D. Milne, E. Frank, and I. Witten, Clustering documents with active learning using wikipedia, In Proceedings of international conference on data mining series, (2008), 839–844.
[10]. A. Huang, D. Milne, E. Frank, and I. Witten, Clustering documents using a wikipedia-based concept representation, In Proceedings of the 13rd PAKDD, (2009), 628–636.
[11]. L. Jing, L. Zhou, M. Ng, and J. Huang, Ontology-based distance measure for text clustering, In Proceedings of the 4th workshop on text mining, the 6th SIAM international conference on data mining, (2006).
[12]. X. Hu, X. Zhang, C. Lu, E. Park, and X. Zhou, Exploiting wikipedia as external knowledge for document clustering, In Proceedings of the 15th ACM SIGKDD,(2009), 389–396.
[13]. E. Gabrilovich, and S.Markovitch, Feature generation for text categorization using word knowledge, In Proceedings of the 19th international joint conference on artificial intelligence, Edinburgh, (2005), 1048–1053.
[14]. E. Gabrilovich, and S. Markovitch, Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge, In Proceedings of the 21st AAAI. Boston, MA, USA, (2006), 1606–1611.
[15]. E. Gabrilovich, and S. Markovitch, Computing semantic relatedness using wikipedia-based explicit semantic analysis, In Proceedings of the 20th IJCAI , (2007), 1606–1611.
[16]. S. Banerjee, K. Ramanathan, and A. Gupta, Clustering short texts using wikipedia, In Proceedings of the 30th ACM SIGIR, (2007), 787–788.
[17]. P. Wang, and C. Domeniconi, Building semantic kernels for text classification using wikipedia, In Proceedings of the 14th ACM SIGKDD. New York, NY, USA, (2008), 713–721.
[18]. Michael Pucher. Performance evaluation of WordNet-based semantic relatedness measures for word prediction in conversational speech, In IWCS 6, Sixth International Workshop on Computational Semantics, Tilburg, Netherlands, (2005).
[19]. G. Demetriou, E. Atwell, and C. Souter, Using lexical semantic knowledge from machine readable dictionaries for domain independent language modeling, In Proc. of LREC 2000, 2nd International Conference on Language Resources and Evaluation, (2000).
[20]. D. Ahn, V. Jijkoun, G. Mishne, K. M¨uller, M. de Rijke and S. Schlobach, Using Wikipedia at the TREC QA track, In Proc. of TREC-13, (2004).
[21]. S. Banerjee, S. and T. Pedersen, Extended gloss overlap as a measure of semantic relatedness, In Proc. of IJCAI-03, (2003), 805. 810.