A Novel Text Categorization Approach based on K-means and Support Vector Machine

Rajesh Malviya
Samrat Ashok Technological Institute
Department of Information Technology,
Vidisha, M.P., India

Pranita Jain
Samrat Ashok Technological Institute
Department of Information Technology,
Vidisha, M.P., India

ABSTRACT

Continuous expansion of digital libraries and online news, the huge amount of text documents is existing on the web. Consequently the need is to organize them. Text Categorization is an active analysis field can be used for organizing text document. Text categorization is the process of assigning documents with predefined categories that are associated with their contented.

CAWP algorithm is designed for Text Categorization. But this algorithm does not present the best results for large datasets. K-means Clustering with Support Vector Machine approach is used to enhance the results. K-means group the data into a number of clusters follow which it uses as training samples for Support Vector Machine in each cluster to divide the new sample data efficiently. The experiment performed on 20Newsgroups dataset, K-means with SVM provides better results than CAWP algorithm in terms of F-measure.

Keywords

Classification, K-means, SVM, Document Categorization, Text mining

1. INTRODUCTION

Recent years have witnessed a massive enlargement in the quantity of textual information existing, both on the World Wide Web as well as in institutional document repositories. In this situation, text mining has happened to extremely operable, give rise to an age where massive amounts of textual information are able to be analyzed, accessed and processed in a fraction of a second. The allowances of text mining goes well outside search and have yielded innovation to facilitate people superior understanding and make utilize of the information in document repositories [1]. Several applications of text mining including in the analysis and classification of news reports, email and spam filtering [2] [3]. Methods to analyze large text corpora fall in two major categories, the first method is text classification and another is text clustering.

Text categorization/classification is the process of assigning documents with predefined categories that are associated with their contented. The text Categorization difficulty has been drawn by means of the pattern-classification process, where documents are treated as numerical vectors and ordinary classifiers (e.g., Decision Tree and Naive Bayes) are applied [5]. This kind of depiction is named Vector Space Model [6]. The Vector Space Model assumes a document as an N-dimensional space and documents so as to close in that space and they are interrelated to each other [7]. Bag-of-words [8] depiction form is used with the variety of instances of Vector Space Model. In the representation of bag- of- word, it is assumed that the content of a document could be determined by the set of terms it has. Within the vocabulary area documents can be represented as points, i.e., a document is represented by a numerical vector of length equal to the numeral of different terms within the vocabulary (the set of all different terms within the document collection). The vector representation specifies how important the consequent terms are describing for the semantics or the content of the document [9]. The bag-of-word is essentially used meant for document illustration in each text categorization and information retrieval. A significant part of the text categorization systems using the bag-of-word representation, hence well-known as term-weighting scheme so as to is responsible for deciding, although the relevant term is for describing the content of a document [10, 11].

The importance of a word in a document is given by the numeral of times that word occurs in the document. Term-weighting methods are termed frequency, where the frequency of occurrence of a word in the document is the weight value of that word. Although, as a result of curing statistical data from the original document provides simplicity of vector space model. It is not simple to do clustering, classification directly in the space, which is exceptionally sparse and high dimensional. Dimension reduction methods are capable of reducing the dimensionality thus to make possible the data can be handled by existing approaches more simply [12]. To solve the problems of text mining there are various techniques presented for retrieval of relevant information. In text mining included information extraction, text categorization, text document analyzed on the basis of term [12], phrase [13], the concept [14] and pattern.

During the long-ago, a variety of clustering and classification methods have been applied for document categorization. Based on the properties of those techniques, general clustering methods are grouped as partitioning clustering and hierarchical clustering, and Naive Bayes and decision tree are examples of classification methods, these are the basic and simple technique for document classification. Documents divided into a numeral of predefined ranges of clusters in partitioning clustering with no hierarchical structure; while in hierarchical clustering documents are grouped into nested partitions of an order [15]. The vector space model is used to represent documents, in the large collected works of document clustering methods, for example k-means [16], clustering around weighted prototype [17], where every document treated as “a bag of words”. To identify the relationship between two vectors, various types of similarity measures have been planned. Euclidean distance is one of them similarity measure which is taken from the Euclidean geometry field [10]. Clustering around weighted prototype for document analysis and categorization is based on the Similarity measure approach. In this approach, every object is assigned by some value of numeric weights which represents the cluster. In this technique numerous weights are assigned for more than one object to every cluster. An object is, the more representative within the corresponding cluster when the larger weight is assigned to that object [17]. This research presents k-means clustering and support vector machine classification based approach.

K-means is simplest unsupervised learning algorithms that resolve the known clustering problem. K-means is applied to a variety of fields to search out approximate best solution in an efficient way. A certain number of clusters (suppose k clusters) fixed a priori in simple and easy manner to categorize a given data throughout [16]. The main thought is to define centroids, one for every cluster. Support vector machine (SVM) is a dominant supervised learning algorithm based on the principal of maximal margin bound. The high generalization ability of the algorithm makes it most suited for high dimensional data such as text. It has been acknowledged as one of the most booming classification algorithms for a lot of applications including text classification [19]. Cluster centers, which generated by the K-means clustering are assumed as a training sample for support vector machine classifier to categorize the new data available. The support vector machine is wildly used in text categorization, because of the High generalization ability for large documents [30].

2. LITERATURE SURVEY

Text Categorization has been examined for a long time. Text or document categorization can be done by using classification or clustering technique. The process of text classification has been expansively intentional and quick improvement seems in this area. K-NN [39] is based on a similarity or distance function for a pair of objects, for example Cosine similarity measure or the Euclidean distance. Naïve bias method is the type of module class [5] underneath well-known priori probability and class conditional probability. The basic idea of this method is to estimate the probability that document D belongs to class C. Another kind of text classification method is decision tree, since it consist of internal node these are labelled by the term, branches are labeled by weight for each term, and class labels are represented by corresponding leaf node [5]. Sequences of nested partitions are created for given data (objects) in Hierarchical clustering [15]. To compute the nearness of two clusters, complete linkage, single linkage and group average linkage are the three approaches are commonly used in hierarchical clustering. Experiments demonstrate that partitioning approaches are superior to hierarchical ones designed for lesser computational cost and improved class of clusters [20]. K-medoids (PAM) [21] is a standard proximity-based approach, which produces k partitions of the data, since every cluster represent each object belongs to that cluster. Based on multiple representative objects that are being proficient to confine the cluster structure, experts create clustering approaches. Clustering using representative, this idea is taken in the hierarchical clustering approach [22], where a number of representative objects prefer to be well divided in order to confine rich cluster form. In [23] different numbers of representative objects for a variety of clusters is proposed, based on cluster density. A multi-representational approach based on density is planned in [25]. Fuzzy clustering with weighted prototype [24], here every cluster is characterized by a variety of weighted medoids. On Partitioned Fuzzy Clustering, the weights as well as the variety of representative objects in every cluster are determined based on the nature of the dataset. Efficiency and scalability are two vital factors to applications with huge scaled data. Another way of solving the problem of text categorization is using both classification and clustering criteria, for instance in paper [40], suggest a large margin classifier that is Clustered Support Vector Machine (CSVM). Here initially data is divided into the numerous clusters by K-means, along with for each cluster, train the support vector machine.

3. TEXT CATEGORIZATION

3.1 Text Document Preprocessing

Before the collections of documents are used for analysis some preprocessing tasks are typically performed. The commonly used tasks are Stemming, Stopword removal, handling of Digits, Hyphens Cases of the Letters and Punctuations.

3.1.1 Stemming

A word has a number of grammatical forms depending on a perspective that it is used in lots of languages. Like to in English, verbs have ground form and verbs as in past tense is different from the present tense, plural forms of nouns. The major reason of steaming is to decrease diverse grammatical word, forms of a word for its adverb, verb, adjective and noun and the rest, in its origin form. Stemming is a procedure that produces stems or roots. A stem is a left part after removing its suffixes and prefixes from the word. For example "walking", "walks" and "walker" is reduced to "walk".

3.1.2 Stopword Removal

Articles, Prepositions, Conjunctions and some pronouns are Stopword. Stopword do not represent any content of the document. Stopword (like a, about, an, are, as at be, from, when, what…….) occurs repeatedly, but do not signify any content of documents. Stopword is less effective in document representation. These words should be removed before documents are stored and indexed. Stopword in the query are also removed before retrieval is performed.

3.1.3 Other pre-processing method for Text

Hyphens:

Breaking hyphens are usually applied to deal by means of inconsistency of usage. There are two types of removal, each hyphen is replaced by a space or each hyphen is simply removing without leaving a space so that "state-of-the-art" may be replaced by "state of the art" or "stateoftheart".

Case of Letters:

All the letters are typically changed to either the upper case or lower case.

Digits: Numbers and terms that include digits are removed for example-dates, times.

Punctuation Marks:

Punctuation is able to be dealt with similarly like hyphens.

3.2 Feature Selection

The goal of the feature selection methods is to approximate the degree of relevance of a document with a score computed based on information such as the frequency of words in the document and the whole collection. In text categorization, features are usually words/terms from a document. Choosing an appropriate feature selection technique for text categorization can be critical because of the large number of features usually present in text documents. Feature selection is extremely important task in document categorization. Feature selection is method of selecting relevant features from the large collection of document set. Commonly used feature selections techniques are document frequency, term frequency-inverse document frequency and term variance scheme are used to choose most important and relevant feature from the large collection of document set.

3.2.1 Document Frequency

Document frequency [26] is used for estimating the efficiency of text documents. Document frequency is the technique of finding out the number of documents in which word/term occurs. Frequent word/terms are more helpful than non frequent words/terms. Document frequency is the quantity of documents through which f_j appears. Document frequency of f_j is specified by the following formula:

$$DF(f_j) = n(f_{ij} > 0) \qquad (1)$$
$$(i = 1,2,....n;\ j = 1,2,....m)$$

$n(f_{ij} > 0)$ is the total number of documents in which f_{ij} is appears.

3.2.2 Variance

Document ranking methods use to rank all documents in the order of relevance. How important a feature is in text document set, defined by manipulative variance of each term. Variance [26] ranks the feature by defining variance of every feature f_j in document term matrix. Variance is specified by the following formula:

$$\text{var}(f_j) = \frac{1}{n}\sum_{i=1}^{n}(f_{ij} - \overline{f}_j)^2 \qquad (2)$$

$$where, \overline{f_j} = \frac{1}{n}(\Sigma_i\, f_{ij})$$

A distinct feature gets high variance score. Variance score are easy feature selection process used for selecting the feature [27].

3.2.3 Vector space model

The main idea consists of representing a document as a vector, in particular as a bag of words. This set contains only the words that belong to the document and their frequency. This means that a document is represented by the words that it contains. In this representation, punctuation is ignored, and a sentence is broken into elementary elements (words) losing the order and the grammar information. These two observations are crucial, because they show that it is impossible to reconstruct the original document given its bag of words; it means that the mapping is not one to one. Consider a corpus as a set of documents, and a dictionary as the set of words that appear into the corpus. This can view a document as a bag of terms or bag of words. This bag can be seen as a vector, where each component is associated with one term from the dictionary. A corpus of ℓ documents can be represented as a document-term matrix whose rows are indexed by the documents and whose columns are indexed by the terms [34]. Each entry at position (i, j) is the term frequency of the term *t* in document *i*.

Term Frequency Scheme

This is the generally well-known weight scheme. The simplest choice to use the raw frequency of a term in a document, means the number of times that term t occurs in document d is said to be term frequency. If we denote the raw frequency of *t* by $tf(i, j)$, then the simple *tf* scheme is tf(t,d) = f(t,d). During this methodology, the weight of term in document is the count of appearance in document.

Let *N* is the total amount of documents within the system or the collection and df_i be the number of documents within which term t_i found at least once. Let f_{ij} is raw frequency count of word t_i in document d_j. And $|v|$ be vocabulary size of set

$$tf_{ij} = \frac{f_{ij}}{\max\left\{ f_{1j,}\, f_{2j} ...., f_{|v|_j} \right\}} \qquad (3)$$

TF-IDF Scheme

The measure of how much information the word provides, for the document is called the inverse document frequency. It defines the importance of term in the whole document collection. That is, whether the term is common or rare in all documents. It is the logarithmically scaled fraction of the documents that contain the word, obtain by dividing the total number of documents by the number of documents containing the term, and then after taking the logarithm of that quotient. Wherever TF is term frequency and IDF is inverse document frequency [33]. Inverse document Frequency of term t_i is given by-

$$idf_i = \log\frac{N}{df_i} \qquad (4)$$

TF-IDF term weight is given by-

$$w_{ij} = tf_{ij} \times idf_j \qquad (5)$$

3.3 Similarity Measure

The similarity between two vectors is considered by the amount of information needed to state the commonality between those two vectors and the information needed to fully describe the vectors. A similarity measure technique finds the commonality of vectors. Sometimes it is also needed to measure the differences between two vectors. Commonly used similarity measure techniques are Jaccard coefficient [27], Cosine similarity [27], and Euclidean distance [18].

4. CLUSTERING AROUND WEIGHTED PROTOTYPE

In this section present the details of the proximity based clustering approach clustering around weighted prototype.

In the dataset with n objects $x = \{x_1, x_2,....x_n\}$ and S is the similarity matrix, where each entry $S_{ij} \in S$ measures the similarity between x_i andx_j, the target is to cluster these n objects into non-overlapping clusters. Weight w_{cj} a continuous variable; to derive the solution of weight w_{cj} by locally maximizing objective function Lagrange multiplier is used. Weight w_{cj} is describes as follows:

$$w_{cj} = \frac{1}{n} + \frac{1}{T}\left[\sum_{x_i \in A_c} s_{ij} - \frac{1}{n}\sum_{q=1}^{n}\sum_{x_i \in A_c} s_{iq}\right] \qquad (6)$$

After driving rule of weight for the objects, then it is needed to define the cluster for each object. The way is to assign each object to the closest cluster. The closeness between object and cluster can be given as:

$$sim(x_i, A_c) = \sum_{r_h \in R_c} v_{ch} sim(x_i, r_h) = \sum_{j-1}^{n} w_{cj} s_{ij} \qquad (7)$$

Thus, x_i is assigned to cluster A_c where

$$c = \arg \max_{f=\{1,2,..k\}} \sum w_{fj} s_{ij} \qquad (8)$$

The alternating optimization is extensively used to obtain local solution of objective-based clustering. Subsequent this process, successively update the cluster assignment and delegate weights base on every other. The representative objects are adjusted from end to end the updating of weights base on the current partitions, as well as the updated representative objects of every cluster in turn are use to form new clusters. Since discuss previously, when $T \rightarrow 0$, CAWP decrease to one object representation for every cluster; while $T \rightarrow \infty$ make all objects uniformly represent every cluster.

As shown in clustering around weighted prototype, rather than fixing the parameter T to a fixed value, Clustering around weighted prototype algorithm starts with a large T_0 and step by step decreases it once the method continues. This is analogy to the annealing procedure, where parameter T will be treated as the temperature. Once the temperature is high, for example T is large, the allocation of weight is close to random and therefore representative objects are search in large space; once T is small, the distribution of weight becomes stable and therefore searches the new delegate objects only from the neighborhood of this one. Such a cooling method allows clustering around weighted prototype to avoid a number of the local optimums and is additional likely to produce better result [17].

Algorithm

Input: Similarity matrix $S_{n \times n}$, the number of cluster k, parameterT_0, T_f,

M = Maximum number of iteration

Output: Cluster to object Weight matrix $W_{k \times n}$.

Steps:

```
I.    $T = T_0$, Initial partition;
II.   Repeat
III.  $\{A_c^{t+1}\}_{c=1}^{k}$
IV.   $T = T_0 \times \left( T_f / T_0 \right)^{t/M}$;
V.    $t = t + 1$.
VI.   Until $t > M, or\ T < T_f$
```

5. PROPOSED ALGORITHM

In this experiment, the text document is too large. To solve the problem of text categorization, there are various techniques are available. In this study firstly applies the partitioning k-means clustering method to generate different cluster centers, assuming cluster centroids as the training sample then applied the support vector machine to categorize text document [29].

The K-means algorithm is the most excellent and simplest clustering algorithms. In the k-means algorithm takes k is the input parameter, and then it divides the n objects into k clusters with the intention that the resultant cluster similarity is high. The Similarity of Cluster is measured by calculating the mean value of the objects within a cluster, so that is known to be the cluster centroids. K-means [28] seeks an optimal partition of the data by minimizing the sum-of-squared-error criterion with an iterative optimization procedure. Usually, the square-error method is used, as defined as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2 \qquad (9)$$

Where,

E is the sum of the square error for the entire objects

p is the point in space signifying a given object and

m_i is the mean of cluster C_i.

The basic clustering procedure of *K*-means is summarized as follows:

Step 1: Randomly choose *k* objects from *D* as the preliminary
Cluster centers;

Step 2: Repeat

Step 3: Reassign every object to the cluster to which the
Object is the mainly similar, base on the mean value
Of the objects within the cluster;

Step 4: Update the cluster means, i.e., compute the mean
Value of the objects for each one cluster;

Step 5: Until no change;

Initially, Centroids are generated via a k-means algorithm are taken as the training samples for classification tasks. Now on the basis of this training data support vector machine is used for text/document categorization for the new test samples [31].

Support Vector Machine (SVM) [38] is a powerful supervised learning algorithm based on the principal of maximal margin bound. The high generalization ability of the algorithm makes it mostly suited for high dimensional data such as text. It has been predictable as one of the most booming classification algorithms for numerous applications as well as text classification. In general, for multiclass classification the most frequent method has been to construct one-versus-rest classifiers (typically known to as “one-versus-all” or OVA classification) where each one category is partitioned out and every one of the other categories are combined and to decide the class which classify the test data with greatest margin. It creates *n* binary classes from an *m* classes. Here in learning step of the classifiers, considering the patterns from the particular class as positives and all other examples as negatives for whole training data. The objective function for support vector machine classifier is given by:

$$\min_{w_m \in \mathrm{H}, \xi \in \Re^l} \frac{1}{2} \sum_{m=1}^{k} w_m^T w_m + C \sum_{i=1}^{l} \xi_i \qquad (10)$$

Subject to constraints,

$$w_{y_i}^T \varphi(X_i) - w_t^T \varphi(X_i) \geq 1 - \delta_{y_i,t} - \xi_i,$$

$$\xi_i \geq 0, i = 1,..., l, t \in \{1,..., k\},$$

Where δ_{ij} is defined as 1 for $i=j$ and as 0 otherwise.

The resulting function is:

$$\arg\max_m f_m(x) = \arg\max_x w_m^T \varphi(x) \qquad (11)$$

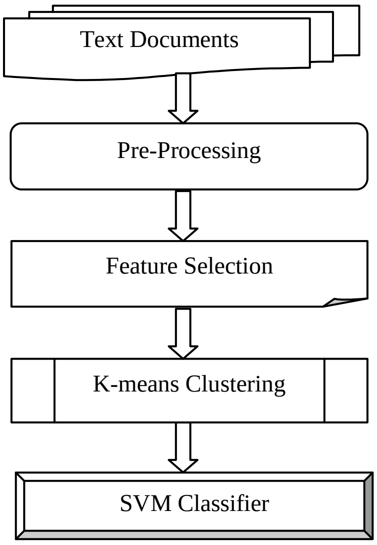

Figure 1: Diagram of Proposed Approach

6. EXPERIMENTAL RESULTS

In this section, suggest the task of document categorization to calculate the Precision, Recall and F-measure. For evaluation, run the K-means with SVM and similarity based technique clustering around weighted prototype. To represent the document corpus vector model is used. The similarity matrix is then generated as input data.

6.1 Datasets Description

The datasets are extracted from 20Newsgroup benchmark. D1, D2 and D3 are the subsets extracted from the 20Newsgroups [37]. 25 features are taken in D1, D2 and D3 those contains around 500, 1000 and 1500 documents respectively from each of five categories comp.graphics, rec.motorcycles, rec.sports.baseball, sci.space, and talk.politics.mideast. Calculate weight for each word by means of term frequency–inverse document frequency weight, and to calculate the similarity between documents Euclidean distance similarity is used.

6.2 Algorithms and Evaluation

The algorithms CAWP [17]: Clustering around weighted prototype algorithm and K-means through Support Vector Machine is run to evaluate the results.

To calculate the quality of the algorithms, precision recall and F-measure [36] is used, produced by Clustering around weighted prototype and K-means through SVM. Precision, Recall and F-measure compares the clusters created by algorithm and taking values in the range of [0, 1].

$$F - measure = \frac{2 \times P \times R}{P + R} \qquad (12)$$

Where *P and R* are defined as:

$$P(precision) = \frac{TP}{TP + FP} \qquad (13)$$

$$R(recall) = \frac{TP}{TP + FN} \qquad (14)$$

Table-1, Table-2 and Table-3 shows the results of Clustering around weighted prototype and K-means with SVM method in terms of Precision, Recall and F-measure respectively. The difference can be seen in Figure-2, Figure-3 and Figure-4 by observing that K-means with SVM method are considerably superior to Clustering around weighted prototype.

Table-1 Result of Precision for 20Newsgroups dataset

Dataset	Precision	
	CAWP	Proposed Approach
D1	0.7764	0.4646
D2	0.7576	0.4659
D3	0.7645	0.6258

Table-2 Result of Recall for 20Newsgroups dataset

Dataset	Recall	
	CAWP	Proposed Approach
D1	0.3523	0.3810
D2	0.2905	0.5273
D3	0.3613	0.5510

Table-3 Result of F-measure for 20Newsgroups dataset

Dataset	F-measure	
	CAWP	Proposed Approach
D1	0.3974	0.4187
D2	0.3011	0.4947
D3	0.4932	0.5860

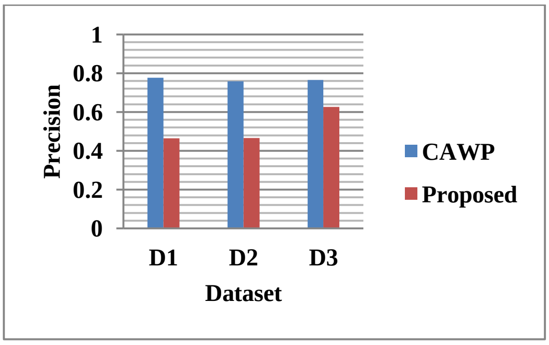

Figure 2: Precision of CAWP and Proposed Approach for 20 Newsgroups dataset

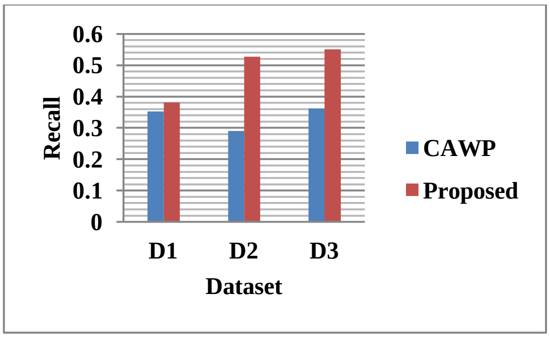

Figure 3: Recall of CAWP and Proposed Approach for 20 Newsgroups dataset

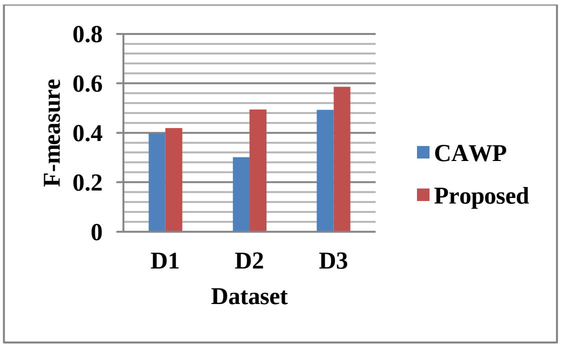

Figure 4: F-measure of CAWP and Proposed Approach for 20 Newsgroups dataset

7. CONCLUSION

Recent years, the development of online digital libraries and online news increases availability of documents in digital form. The task of organizing these documents within its exact categories and to expose meaningful information from large amount of documents is not so easy to. Text categorization is the method of attaching documents with predefined categories that are associated with their contented. Text Categorization is a vital and active analysis field intended for the motive that the huge number of documents existing and the requirement is organizing them.

The categorization of textual document Clustering around Weighted Prototype algorithm is used, but this algorithm does not provide better results for large dataset. This experiment, evaluates the performance of Clustering around Weighted Prototype and K-means with Support Vector machine based approach for text categorization.

Precision, Recall and F-Measure performance measurements are used to compare the result of K-means and SVM based text categorization with Clustering around Weighted Prototype method. The techniques are tested on 20newsgroups datasets and it found that K-means and SVM based text classifier provides efficient results as compared with Clustering around Weighted Prototype. The results are presented in experiment shows that F-measure and Recall of K-means and SVM based text classification approach is superior than clustering around weighted prototype.

There are further improvements can be done on the performance of this K-means and SVM based text classifier. The possible evolutionary techniques that may be used to improve results of K-means and SVM based text classifier and boost the work existing in this investigation. Newly developed evolutionary methods like PSO and cohort Intelligence may be integrated to find better solutions of the objective function. For text representations by NLP techniques to further get better the performance of document categorization.

8. 8. REFERENCES

[1] Li, Y. H., & Jain, A. K. (1998). Classification of text documents. *The Computer Journal*, *41*(8), 537-546.

[2] Cormack, G. V., Smucker, M. D., & Clarke, C. L. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. Information retrieval, 14(5), 441-465.

[3] Kallipolitis, L., Karpis, V., & Karali, I. (2012). Semantic search in the world news domain using automatically extracted metadata files. Knowledge-Based Systems, 27, 38-50.

[4] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

[5] Jun, S., Park, S. S., & Jang, D. S. (2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness.Expert Systems with Applications, 41(7), 3204-3212.

[6] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

[7] Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37(1), 141-188.

[8] Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press.

[9] Altınçay, H., & Erenel, Z. (2010). Analytical evaluation of term weighting schemes for text categorization. Pattern Recognition Letters, 31(11), 1310-1323.

[10] Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(4), 721-735.

[11] Debole, F., & Sebastiani, F. (2004). Supervised term weighting for automated text categorization. In Text mining and its applications (pp. 81-97). Springer Berlin Heidelberg.

[12] Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools, 13(01), 157-169.

[13] Caropreso, M. F., Matwin, S., & Sebastiani, F. (2000). Statistical phrases in automated text categorization. Centre National de la Recherche Scientifique, Paris, France.

[14] Shehata, S., Karray, F., & Kamel, M. (2007, August). A concept-based model for enhancing text categorization. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 629-637). ACM.

[15] Xu, R., & Wunsch, D. (2008). Clustering (Vol. 10). John Wiley & Sons.

[16] Zhong, S. (2005, August). Efficient online spherical k-means clustering. InNeural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on (Vol. 5, pp. 3180-3185). IEEE.

[17] Mei, J. P., & Chen, L. (2014). Proximity-based k-partitions clustering with ranking for document categorization and analysis. Expert Systems with Applications, 41(16), 7095-7105.

[18] Schoenharl, T. W., & Madey, G. (2008). Evaluation of measurement techniques for the validation of agent-based simulations against streaming data. InComputational Science–ICCS 2008 (pp. 6-15). Springer Berlin Heidelberg.

[19] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features (pp. 137-142). Springer Berlin Heidelberg.

[20] Zhao, Y., Karypis, G., & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. Data mining and knowledge discovery, 10(2), 141-168.

[21] Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

[22] Guha, S., Rastogi, R., & Shim, K. (2001). Cure: an efficient clustering algorithm for large databases. Information Systems, 26(1), 35-58.

[23] Bellec, J. H., & Kechadi, T. M. (2007, November). Cufres: clustering using fuzzy representative eventsselection for the fault recognition problem intelecommunication networks. In Proceedings of the ACM first Ph. D. workshop in CIKM (pp. 55-62). ACM.

[24] Mei, J. P., & Chen, L. (2010). Fuzzy clustering with weighted medoids for relational data. Pattern Recognition, 43(5), 1964-1974.

[25] Halkidi, M., & Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. Pattern Recognition Letters, 29(6), 773-786.

[26] Liu, L., Kang, J., Yu, J., & Wang, Z. (2005, November). A comparative study on unsupervised feature selection methods for text clustering. In Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on (pp. 597-601). IEEE.

[27] González, C. G., Bonventi Jr, W., & Rodrigues, A. V. (2008). Density of closed balls in real-valued and autometrized boolean spaces for clustering applications. In Advances in Artificial Intelligence-SBIA 2008 (pp. 8-22). Springer Berlin Heidelberg.

[28] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(7), 881-892.

[29] Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier.

[30] Tan, S. (2008). An improved centroid classifier for text categorization. Expert Systems with Applications, 35(1), 279-285.

[31] Ahuja, Y., & Yadav, S. K. (2012). Multiclass classification and support vector machine. Global Journal of Computer Science and Technology Interdisciplinary,12(11).

[32] Lin, J., Li, X., & Jiao, Y. (2010, March). Text Categorization Research Based on Cluster Idea. In Education Technology and Computer Science (ETCS), 2010 Second International Workshop on (Vol. 1, pp. 483-486). IEEE.

[33] Wang, Z., & Qian, X. (2008, December). Text categorization based on LDA and SVM. In Computer Science and Software Engineering, 2008 International Conference on (Vol. 1, pp. 674-677). IEEE.

[34] Srivastava, A. N., & Sahami, M. (Eds.). (2009). Text mining: Classification, clustering, and applications. CRC Press.

[35] Berry, M. W., & Kogan, J. (Eds.). (2010). Text mining: applications and theory. John Wiley & Sons.

[36] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management,45(4), 427-437.

[37] Lang, K. (1995, July). Newsweeder: Learning to filter netnews. In Proceedings of the 12th international conference on machine learning (pp. 331-339).

[38] Wang, Z., & Xue, X. (2014). Multi-Class Support Vector Machine. In Support Vector Machines Applications (pp. 23-48). Springer International Publishing.

[39] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (pp. 986-996). Springer Berlin Heidelberg.

[40] Gu, Q., & Han, J. (2013). Clustered support vector machines. In proceedings of the sixteenth international conference on artificial intelligence and statistics (pp. 307-315).