Semi Supervised Document Classification Model Using Artificial Neural Networks

Dr.M.Karthikeyan
Programmer (Senior Scale)
Department of Computer Science And Engineering, Annamalai University, Annamalai Nagar, India
Mobile Number – (91)- 9443665646

Abstract: *Automatic document classification is of paramount importance to knowledge management in the information age. Document classification is a kind of text data mining and organization technique that automatically groups related documents into clusters. Most of the common techniques in document classification are based on the statistical analysis of a term, either word or phrase. Statistical analysis of a term frequency captures the importance of the term within the document only. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term. To solve this problem the proposed system concentrates on an interactive text clustering methodology, semi supervised document classification method using neural networks. There are two main phases in the proposed method: Pre-processing phase and Classification phase. In the pre-processing phase, distinct words are identified and their frequency of occurrences in the document corpus is calculated. These discovered distinct words with their frequency of occurrences, form a document vector. In the classification phase, Back propagation algorithm is used for document classification by using the feature vector of distinct words. The proposed method evaluates the system efficiency by implementing and testing the clustering results with Dbscan and K-means clustering algorithms. Experiment shows that the proposed document clustering method performs with an average efficiency of 92% for various document categories..*

Keywords — *Artificial Neural Network (ANN), Self Organizing Map(SOM), Back Propagation Networks (BPN), Term frequency, Tokenization, Structural filtering.*

I. INTRODUCTION

Document classification can be defined as the task of learning methods for categorizing collections of electronic documents into their automatically annotated classes, based on its contents. For several decades now, document classification in the form of text classification systems have been widely implemented in numerous applications such as spam filtering, e-mail categorization, formation of knowledge repositories and ontology mapping. An increasing number of statistical approaches have been developed for document classification including k-nearest neighbour classification, naïve bayes classification, support vector machines, maximum entropy and decision tree induction. Each one of the document classification schemes mentioned previously has unique properties. The decision tree induction algorithm and rule induction algorithm are simple to understand and interpret after a brief explanation. However these algorithms do not work well when the number of distinguishing features is large. The k-nearest-neighbor algorithm is easy to implement and has show its effectiveness in a variety of problem domains. However the main drawback of k-nearest-neighbor algorithm is that it is computationally intensive, especially when the size of the training set grows. Support vector machines can be used as a discriminative document classifier, and these have been shown to be more accurate in classification tasks. The good generalization property of the SVM is due to the implementation of structural risk minimization which entails finding a hyper-plane which guarantees the lowest classification error. An ability to learn which is independent of the dimensionality of the feature space is also an advantage of the SVM. However, the usage of SVMs in many real world applications is relatively complex due to its convoluted training and categorizing algorithms as compared to the naïve bayes classifier.

Among these approaches, the naïve bayes text classifier has been widely used because of simplicity in both the training and classifying stage although this generative method has been reported less accurate than discriminative methods such as SVM. However, some researchers have proven that the naïve bayes classification approach provides an intuitively simple text generation model and performs surprisingly well in many other domains, under specific ideal conditions. A naive bayes classifier is a simple probabilistic classifier based on bayes theorem with strong independence assumptions, but this assumption severely limits its applicability. In real life applications, the probability values associated with an event are seldom ‘‘independent”. For example, even tossing a coin will not have the expected 50:50 chance of a result being either ‘‘heads” or ‘‘tails” due to factors which are associated with machining the coin, the different

surface textures, different environments and different ways and methods used to toss the coin among other things. If we are lucky, these factors even out over time, if we are not, then the naïve bayes formula will misclassify frequently. However, depending on the precise nature of the probability model, naive bayes classifiers can be trained very efficiently and requires a relatively small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Artificial neural network (ANN) is an information processing method inspired by biological nervous systems. An ANN uses interconnected processing nodes computationally linked to solve problems. Neural networks are frequently used for pattern recognition and document classification and learn by using training data to adjust the weights between connecting nodes. Some research has applied artificial neural networks to text classification. The self organizing map (SOM) is a clustering method which clusters data, based on a similarity measure related to the calculation of Euclidean distances. The idea of this principle is to find a winner-takes-all neuron to find the most closely matching case. The SOM was proposed by Kohonen, and is based on the idea that systems can be design to emulate the collective co-operation of the neurons in the human brain. Collectivism can be realized by feedback and thus can also be realized in the network, where many neighbouring neurons react collectively upon being activated by events. If neurons are activated in the learning process, the neighboring neurons are also affected. The network structure is defined by synapses and has a similar total arrangement after a phase of self organization as the input data of the event space. Consequently, the SOM is an established paradigm in AI and cognitive modeling being the basis of unsupervised learning. This unsupervised machine learning method is widely used in data mining, visualization of complex data, image processing, speech recognition, process control, diagnostics in industry and medicine, and natural language processing. Based on these observations, the proposed method concentrates on automatic concept based semi supervised document classification approach which uses back propagation neural network.

The structure of this paper is as follows: Section 2 discusses some related research work regarding document clustering. Section 3 provides system overview and describes how document clustering done by Back Propagation algorithm. The experimental results for the proposed method are presented in section 4. Conclusion and discussion is given in section 5. Section 6 shows the references made.

II. REVIEW OF RELATED LITERATURE

Document clustering is a powerful technique to detect topics and their relations for information browsing, analysis and organization. In recent studies, many new technologies are introduced. Yuen-Hsien Tseng [1] proposed an algorithm for cluster labeling for creating generic titles based on external resources such as wordNet. This method first extracts category-specific terms as cluster descriptors, and then these descriptors are mapped to generic terms based on a hypernym search algorithm. Pei-Yi Hao, Jung-Hsien Chiang, Yi-Kun Tu [2] proposed a novel hierarchical classification method that generalizes support vector machine learning.

Ramiz M. Aliguliyev [3] developed a method to show assignment weight to documents that improves clustering solution because document clustering has been traditionally investigated as a means of improving the performance of search engines by pre-clustering the entire corpus. Linghui Gong, Jianping Zeng, Shiyong Zhang [4] proposed a validity index based method of adaptive feature selection, incorporating with which a new text stream clustering algorithm. Ridvan Saracoglu, Kemal Tutuncu, Novruz Allahverdi [5] developed a method for finding similar documents, that uses predefined fuzzy clusters to extract feature vectors of related documents. Similarity measure is based on these vectors and [6], they proposed a new approach on search for similar documents with multiple categories using fuzzy clustering that uses fuzzy similarity classification method and multiple categories vector method. Shih-Cheng Horng, Feng-Yi Yang, Shieh-Shing Lin [7] proposed a hierarchical fuzzy clustering decision tree for the classification problem with large number of classes and continous attributes. Wei Song, Cheng Hua Li, Soon Cheol Park [8] developed a method that uses genetic algorithm for text clustering based on ontology and evaluating the validity of various semantic measures.

Shady Shehata, Fakhri Karray and Mohamed S. Kamel [9] proposed a new concept based mining model that analyzes terms on the sentence, document and corpus levels. It can effectively discriminate between non important terms with respect to sentence semantics. Hung Chim, Xiaotie Deng [10] developed a method for efficient phrase based document similarity for clustering of documents. They used phrase-based document similarity to compute the pairwise similarities of documents based on suffix. Alexander A. Frolov, Dusan Husek, Pavel Yu.Polyakov [11] introduced a neural-network based algorithm for word clustering. Cheng Hua Li and Soon Cheol Park [12] proposed a new text

classification models based on artificial neural networks and Singular Value Decomposition. The neural networks are trained by Multi-Output Perceptron Learning algorithm and Back-Propagation Neural Network. Jie Ji, Kunita Daichi and Qiangfu [13] developed a customer intention aware system for document analysis. The system starts from an unlabeled document set, give out several cluster results. Tommy W.S. Chow, M.K.M. Rahman [14] proposed a new document retrieval and plagiarism detection system using multilayer self-organizing map. A document is modeled by a rich tree-structured representation, and a SOM based system is used as a computationally effective solution. Zhonghui Feng, Junpeng Bao, Junyi Shen [15] developed a novel dynamic and adaptive SOM algorithm applied to high dimensional large scale text clustering. Dino Isa, Rajprasad Rajkumar, Grham Kendall [16] compares the recognition performance of text and non text images. Hemalatha.M, Sathya Srinivas. D [17] addresses the problems of document mining related with web page clustering and classification. M. Karthikeyan, P. Aruna [18] proposed a new method for clustering based on probability of occurrences of words. Kantu. Vijaya Kumar, Abburi. Venkatesh [19] proposed Multi-Document summarization using phrase context based Indexing and Geometric Model. Mulluri Raghupathi, R. Lakshmi Tulasi [20] proposed a new algorithm which is Hierarchical Filter based Document Clustering.

III. SYSTEM OVERVIEW AND METHODOLOGY

(A). Pre-processing - Probability based topic oriented and semi supervised document clustering method is used in the pre-processing stage. It is defined as follows: Given a set S of N documents and a set T of k topics, proposed system like to partition the documents into k subsets S_1, $S_{2, \ldots}S_k$, each corresponding to one of the topics, such that (i). The documents assigned to each subset are more similar to each other than the documents assigned to different subsets, and (ii). The documents of each subset are more similar to its corresponding topic than the rest of the topics. The functional components of pre-processing stage are depicted in Figure 1. In pre-processing documents are grouped according to the user's need. The main steps includes: (1) Design a multiple-attributes topic structure to represent user's need. (2) Make topic-semantic annotation for each document, and then compute topic-semantic similarity between documents. (3) Compute distinct words probability (4) Group documents based on maximum probability of distinct words. The main objective is to reduce dimensionality of feature vectors. Dimensionality reduction of feature vectors is a hotspot of research in text mining. The dimensionality of document vectors may reach to thousands and even tens of thousands. It results in huge time cost on documents clustering. But in the proposed system distinct words of maximum probability are mapped to attribute of the topic directly. So, dimension is reduced effectively.

Five different categories of documents are used for training purpose as a sample document corpus. They are Business documents, Education documents, Politics documents, Medicine documents and Sports documents. But, it is possible to extend the categories.

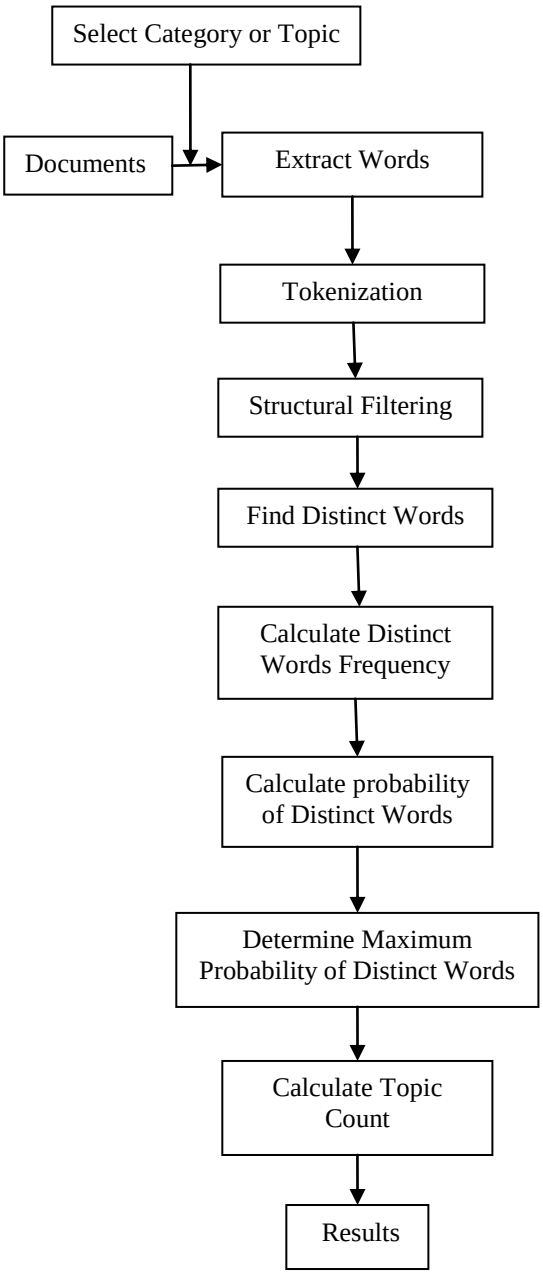

Figure 1: Functional Components of Pre-Processing Stage.

The aim is to derive higher level concepts from the words of the different categories of document corpus in order to populate the knowledge base (database). In the first task, words are extracted from the

training documents and they are matched with other words, or with that already exist in the database. To achieve this goal, a pipeline of analysis that contains two stages is defined. The two stages of task include:
Tokenization - This is the very first linguistic analysis step. It consists of breaking the free text into a sequence of separate words and punctuation symbols (tokens). It input consist of natural language text and the output contains a list of the tokens extracted.
Structural filtering -This stage uses the output of the tokenization, keeps, discovers or discards words according to contextual information. The actual module is a rule compiler which applies filtering based on rules like, words length less than 3 characters and greater than 20 characters etc., Option is given in the proposed system to omit stop words like,"Can","are","has","with","the","they","which"," have",etc. Similarly, words above 15 or 20 characters are also omitted because these words will not be distinct words for document clustering process. Then, the filtered text (words) is called distinct words and these words are stored in the database based on selected category by the user. Then, the probability of occurrence is calculated for distinct words. For example a set S of N documents and a set T of K topics, then the probability of distinct words will be calculated by $\varepsilon W_i / W_j$ Where W_i denotes the number of occurrences of a distinct word in a training document and W_j denotes the total number of distinct words in a training document. Similarly for K topics of set T, the probabilities of occurrences of distinct words are calculated for a set S of N documents.

(B). Classification of Documents Using Back Propagation Algorithm

Back propagation is similar to LMS (least mean squared error) learning algorithm and is based on gradient descent; weights are modified in a direction that corresponds to the negative gradient of an error measure. The choice of everywhere differentiable node functions allows correct application of this method. The major advance of Back propagation over the LMS and perceptron algorithms is in expressing how an error at a higher (or outer) layer of a multiplayer network can be propagated backwards to nodes at lower (or inner) layers of the network; the gradient of these backward propagated error measures (for inner layer nodes) can then be used to determine the desired weight modifications for connections that lead into these hidden nodes. The back propagation algorithm has a major impact on the field of neural networks and has been widely applied to a large number of problems in many disciplines.

Artificial Neuron

The artificial neuron was designed to mimic the first order characteristics of the biological neuron. In essence, set of inputs are applied, each representing the output of another neuron. Each input is multiplied by a corresponding weight, analogous to synaptic strength, and all of the weighted inputs are then summed to determine the activation level of the neuron. The figure 2 shows a model of artificial neuron. By using artificial neuron the Back propagation neural network was framed.

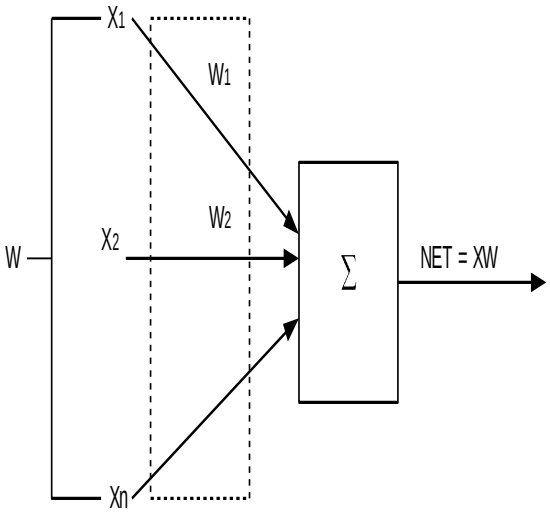

Figure 2: Artificial neuron

$$NET = x_1w_1+x_2w_2+\ldots+x_nw_n$$

Here, a set of inputs labeled x1,x2,…,xn is applied to the artificial neuron. These inputs, collectively referred to as the vector X, corresponds to the signals into the synapses of biological neuron. Each signal is multiplied by an associated weight w1,w2,…wn, before it is applied to the summation block, labeled ∈. Each weight corresponds to the 'strength' of a single biological synaptic connection. The summation block, corresponding roughly to the biological cell body, adds all of the weighted inputs algebraically, producing an output that we call NET. This may be compactly stated in vector notation as follows:

$$NET = XW$$

Back Propagation Algorithm

Forward Pass

1. Apply the inputs.
2. Calculate the NET of each neuron in the hidden layer.
3. Calculate the OUT of each neuron in the hidden layer by substituting,
   $$OUT = 1/(1+e^{-net})$$
4. The output calculated from the hidden layer becomes the input for the output layer.
5. Repeat step 2 and 3 for the output layer.
6. Calculate δ(error) = OUT(1-OUT)(TARGET-OUT) for each neuron in the output layer for each output.

Reverse Pass

1. Adjust the weights of the layer, considering neuron 'q' in the output layer k

$\Delta w\ pq,\ k = \eta\delta\ q,k + OUTpj$
$w\ pq,k(n+1) = w\ pq,k(n) + \Delta pq,k$

2. Adjust the weights in the hidden layer by,
$\delta\ pq = OUT\ Pj\ (1-OUT\ Pj)\ \Sigma\delta\ q,k\ W\ pq,k$
$Wij\ (new) = Wij\ (old) + \eta.\delta\ pq\ Xi$

The sample Neural Network is shown in figure 3 and its NET, OUT and error value calculations are summarized below.

Forward Pass

Net11 = X1*U11+X2*U21+X3*U31
Out11 = $1/1+e^{-Net11}$
Net12 = X1*U12+X2*U22+X3*U32
Out12 = $1/1+e^{-Net12}$
Net13 = X1*U13+X2*U23+X3*U33
Out13 = $1/1+e^{-Net13}$
Net21 = Out11*V11+Out12*V21+Out13*V31
Out21 = $1/1+e^{-Net21}$
Net22 = Out11*V12+Out12*V22+Out13*V32
Out22 = $1/1+e^{-Net22}$
Net31 = Out21*W11+Out22*W21
Out31 = $1/1+e^{-Net31}$
Net32 = Out21*W12+Out22*W22
Out32 = $1/1+e^{-Net32}$
Net33 = Out21*W13+Out22*W23
Out33 = $1/1+e^{-Net33}$
δ31 = Out31(1-Out31)(Target-Out31)
δ32 = Out32(1-Out32)(Target-Out32)
δ33 = Out33(1-Out33)(Target-Out33)

Reverse Pass

W11(new) = W11(old)+ δ31*Out21
W12(new) = W12(old)+ δ32*Out21
W13(new) = W13(old)+ δ33*Out21
W21(new) = W21(old)+ δ31*Out22
W22(new) = W22(old)+ δ32*Out22
W13(new) = W23(old)+ δ33*Out22

δ21 = Out21(1-Out21)(δ31*W11 + δ32*W12 + δ33*W13)
δ32 = Out22(1-Out22)(δ31*W21 + δ32*W22 + δ33*W23)

V11(new) = V11(old)+ η*δ21*Out11
V12(new) = V12(old)+ η*δ22*Out11
V21(new) = V21(old)+ η*δ21*Out12
V22(new) = V22(old)+ η*δ22*Out12
V31(new) = V31(old)+ η*δ21*Out13
V32(new) = V32(old)+ η*δ22*Out13

δ11 = Out11(1-Out11)(δ21*V11)+ δ22*V12
δ12 = Out12(1-Out12)(δ21*V21)+ δ22*V22
δ13 = Out13(1-Out13)(δ21*V31)+ δ22*V32

U11(new) = U11(old)+ η*δ11*X1
U12(new) = U12(old)+ η*δ12*X1
U13(new) = U13(old)+ η*δ13*X1
U21(new) = U21(old)+ η*δ11*X2

U22(new) = U22(old)+ η*δ12*X2
U23(new) = U23(old)+ η*δ13*X2
U31(new) = U31(old)+ η*δ11*X3
U32(new) = U32(old)+ η*δ12*X3
U33(new) = U33(old)+ η*δ13*X3

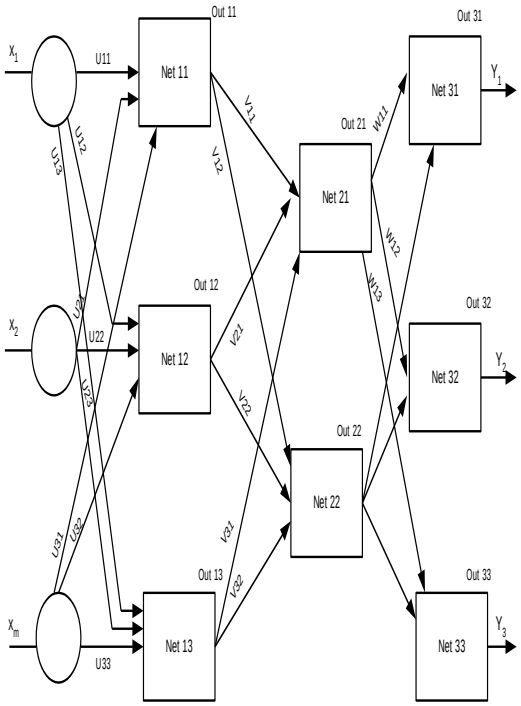

Figure 3: Multi Layer Back propagation Neural Network.

The distinct words with high probability which are identified in pre-processing stage are given as input to the back propagation network. During forward pass NET and OUT values are calculated for each neuron. Since, the proposed system uses semi supervised approach, error values are calculated by comparing the actual output by the desired output. Based on the error values new weight values are calculated. This process was repeated for all layers. For achieving more accuracy the proposed system uses reverse pass. During reverse pass the process was repeated by adjusting the weight values. Forward pass and Reverse pass are executed until the error rate is considerably low. Efficiency is calculated by *Number of documents classified correctly / Total number of documents.*

IV. RESULTS AND DISCUSSIONS

In experiments, totally 1000 documents of five different categories are used for training purpose. The system was trained according to the category selected by the user. The categories include: 1. Business documents, 2. Education documents, 3. Political documents, 4. Medicine documents and 5.Sports documents. Proposed system omits nearly 500 stop words. The stop words in document corpus are removed before distinct words extraction. A partial list of stop words extracted is presented in

table 1. Then, distinct words are extracted and their frequency of occurrence is calculated. Based on the frequency, probability is calculated. Table 2 shows training process, results of Back Propagation algorithm. Totally during the experiments 401914 distinct words are identified. According to the category selected by the user, the probability was calculated. Based on the probability of distinct words clustering is done during the training phase. For testing, 250 documents (50 documents in each category) are used. The accuracy comparison between document categories is shown in figure.4. Table 3 shows the accuracy achieved Back Propagation algorithm for each category of the documents. The same documents are used as test case for Dbscan and K-means clustering algorithms Table 4 shows accuracy achieved by Dbscan algorithm and Table 5 shows the accuracy achieved by the K-means algorithm. When comparing the existing Dbscan algorithm, K-means algorithm, the proposed semi-supervised document clustering model using artificial neural networks yields better results. Figure 4 shows the overall comparison of performance by all the three clustering algorithms.

Table 1 A partial list of stop words

are	as	at	all	after	beyond
around	anyone	again	always	by	besides
can	could	both	but	Have	for
be	how	that	there	each	Last
else	every	few	be	it	me

Table 2 Training process results of Back Propagation Training Algorithm

Sl. No	Category	Number of documents Used for training	Total number of distinct Words extracted
1	Business	200	78127
2	Education	200	81679
3	Politics	200	78010
4	Medicine	200	82342
5	Sports	200	81756

Table 3 Accuracy of Back Propagation Algorithm

Category	Number of documents Used for testing	Number of documents Classified correctly	percentage
Business	50	45	90%
Education	50	46	92%
Politics	50	44	88%
Medicine	50	47	94%
Sports	50	48	96%

Table 4 Accuracy of Dbscan Algorithm

Category	Number of documents Used for testing	Number of documents Classified correctly	percentage
Business	50	42	84%
Education	50	40	80%
Politics	50	39	78%
Medicine	50	38	76%
Sports	50	41	82%

Table 3 Accuracy of K-means Algorithm

Sl. No	Category	Number of documents Used for training	Total number of distinct Words extracted
Business	50	38	76%
Education	50	39	78%
Politics	50	36	72%
Medicine	50	35	70%
Sports	50	38	76%

The Dbscan algorithm performs with an average efficiency of 80% for all categories. K-means algorithm performs with an average efficiency of 74.4% for all categories. The proposed method performs with an average efficiency of 92% for all categories of documents. The results show that the proposed semi supervised document clustering model using artificial neural network outperforms other two existing methods.

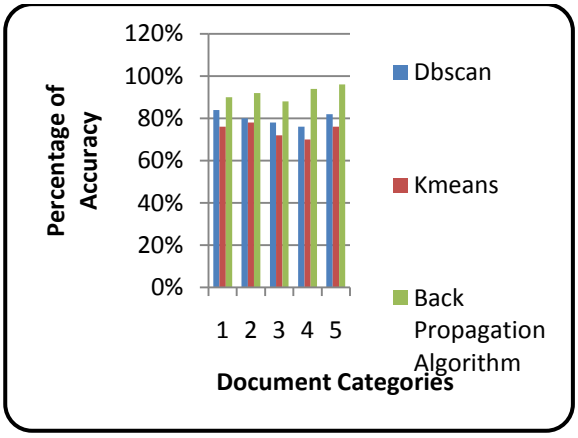

Figure 4 Overall performance comparison

V. CONCLUSIONS

Traditional document clustering are unsupervised learning approaches. Traditional approaches often fail to obtain good clustering solution when users want to group documents according to their need. Focusing on this problem, the proposed method uses the semi supervised document clustering using Back

propagation algorithm to fulfil the user requirement. Further the proposed method was compared and checked with the famous clustering algorithms Dbscan and K-means. In future probability similarity score to include arbitrary functions over words in documents (such as phrases and logical operations) may be implemented. This can be done by expanding the domain of the multi-nominal distributions currently used to compute expected document overlap and similarity measure may be extended to problems in other domains such as video segmentation, audio classification using different estimation techniques as appropriate. Although one thousand documents of various categories used for training purpose in the proposed system, it is scalable according to the user's need. Experiments show that the proposed method is feasible and effective.

VI. REFERENCES

[1] Yuen - Hsien Tseng, Generic title labeling for clustered documents, Expert Systems with Applications, 37(2010) 2247-2254.

[2] Pei-Yi Hao, Jung - Hsien Chiang, Yi – Kun Tu, Hierarchically SVM classification based on support vector clustering method and its application to document categorization, Expert Systems with Applications, 33(2007) 627-635.

[3] Ramiz M. Aliguliyev, Clustering of document collection – A weighting approach, Expert Systems with Applications, 36(2009) 7904-7916.

[4] Linghui Gong, Jianping Zeng, Shiyong Zhang, Text stream clustering algorithm based on adaptive feature selection, Expert Systems with Applications, 38(2011) 1393-1399.

[5] Ridvan Saracoglu, Kemal Tutuncu, Novruz Allahverdi, A fuzzy clustering approach for finding similar documents using a novel similarity measure, Expert Systems with Applications, 33(2007) 600-605.

[6] Ridvan Saracoglu, Kemal Tutuncu, Novruz Allahverdi, A new approach on search for similar documents with multiple categories using fuzzy clustering, Expert Systems with Applications, 34(2008) 2545-2554.

[7] Shih-Cheng Horng, Feng - Yi Yang, Shieh -Shing Lin, Hierarchical fuzzy clustering decision tree for classifying recipes of ion impanter, Expert Systems with Applications, 38(2011) 933-940.

[8] Hung Chim, Xiaotie Deng, Efficient Phrase –Based Document Similarity for Clustering, IEEE Transactions on Knowledge and Data Engineering, Vol 20,No.9(2008).

[9] Shady Shehta, Fakhri Karray, Mohamed S. Kamal, An Efficient Concept-Based Mining Model for Enhancing Text Clustering, IEEE Transactions on Knowledge and Data Engineering, vol. 22, No.10, October 2010.

[10] Hung Chim, Xiaotie Deng, Efficient Phrase –Based Document Similarity for Clustering, IEEE Transactions on Knowledge and Data Engineering, Vol 20,No.9(2008).

[11] Alexander A. Frolov, Dusan Husek, Pavel Yu .Polyakov, Recurrent-Neural – Network Based Boolean Factor Analysis and Its Application to Word Clustering, IEEE Transactions on Neural Networks, Vol 20,No.7(2009).

[12] Cheng Hua Li and Soon Cheol Park, Neural Network for Text Classification Based on Singular Value Decomposition, Seventh International Conference on Computer and Information Technology, 0-7695-2986-6/07, IEEE, 2007.

[13] Jie Ji, Kunita Daichi and Qiangfu, A Customer Intention Aware System for Document Analysis, 978-1-4244-8126-2/10,IEEE, 2010.

[14] Tommy W.S. Chow, M.K.M. Rahman, Multilayer SOM with Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection, IEEE Transactions on Neural Networks, Vol 20, No.9, 2009.

[15] Zhonghui Feng, Junpeng Bao, Junyi Shen, Dynamic and Adaptive Self Organizing Maps applied to High Dimensional Large Scale Text Clustering, 978-1-4244-6055-7/10, IEEE, 2010.

[16] Dino Isa, Rajprasad Rajkumar, Grham Kendall, Document Zone Classification for Technicial Document Images Using Artificial Neural Networks and Support Vector Machines, 978-1-4244-4457-1/09, IEEE, 2009.

[17] Hemalatha.M, Sathya Srinivas. D, Hybrid Neural Network Model for Web Document Clustering, 978-1-4244-4457-1/09, IEEE, 2009.

[18] M. Karthikeyan, P.Aruna, Probability Based document clustering and Image clustering using Content Based Image Retrieval, Applied Soft Computing, Vol 13, 959-966,2013.

[19] Kantu. Vijaya Kumar, Abburi. Venkatesh, Multi-Document summarization using phrase context based Indexing and Geometric Model, International Journal of Computer Trends and Technology (IJCTT) – volume 17 Number 5 Nov 2014.

[20] Mulluri Raghupathi, R. Lakshmi Tulasi, Hierarchical Filter based Document Clustering Algorithm, International Journal of Computer Trends and Technology (IJCTT) – volume 21 Number 1 Nov 2014.