International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.2, Issue.5, Sep.-Oct. 2012 pp-3131-3137 ISSN: 2249-6645

An Overview of Categorization techniques

B. Mahalakshmi[1], Dr. K. Duraiswamy[2]

[1]*Associate Prof., Dept. of Computer Science and Engineering, K. S. Rangasamy College of Technology, Tiruchengode, India*
[2]*Dean (Academic), K. S. Rangasamy College of Technology, Tiruchengode, India*

***Abstract* :** *Categorization is the process in which ideas and objects are recognized, differentiated and understood. Categorization implies that objects are grouped into categories, usually for some specific purpose. A category illuminates a relationship between the subjects and objects of knowledge. The data categorization includes the categorization of text, image, object, voice etc .With the rapid development of the web, large numbers of electronic documents are available on the Internet. Text categorization becomes a key technology to deal with and organize large numbers of documents. Text representation is an important process to perform text categorization. A major problem of text representation is the high dimensionality of the feature space. The feature space with a large number of terms is not only unsuitable for neural networks but also easily to cause the over fitting problem. Text categorization is the assignment of natural language documents to one or more predefined categories based on their semantic content is an important component in many information organization and management tasks. This paper discusses various categorization techniques, tools and their applications in different fields.*

Keywords- *Clustering, Neural networks, Latent Semantic Indexing, Self-Organizing map.*

I. Introduction

Automatic text categorization is an important application and research topic for the inception of digital documents. Text categorization [1] is a necessity due to the very large amount of text documents that humans have to deal with daily. A text categorization system can be used in indexing documents to assist information retrieval tasks as well as in classifying e-mails, memos or web pages in a yahoo-like manner.

The text classification task can be defined as assigning category labels to new documents based on the knowledge gained in a classification system at the training stage. In the training phase, given a set of documents with class labels attached and a classification system is built using a learning method, machine learning communities.

Text classification [4] tasks can be divided into two sorts: supervised document classification where some external mechanism provides information on the correct classification for documents, and unsupervised document classification, where the classification must be done entirely without reference to external information. There is also a semi-supervised document classification, where some documents are labeled by the external mechanism.

Text categorization [2] is the problem of automatically assigning predefined categories to free text documents. While more and more textual information is available online, effective retrieval is difficult without indexing and summarization of document content. Document categorization is one solution to this problem. A growing number of statistical classification methods and machine learning techniques has been applied to text categorization including Neural Networks, Naïve Bayes classifier approaches, Decision Tree, Nearest neighbor classification, Latent semantic indexing, Support vector machines, Concept Mining, Rough set based classifier, Soft set based classifier[3].

Document classification techniques include:

- Back propagation Neural Network
- Latent semantic indexing
- Support vector machines
- Decision trees
- Naive Bayes classifier
- Self-Organizing Map
- Genetic Algorithm.

II. Back Propagation Neural Network

The back-propagation neural network [5] is used for training multi-layer feed-forward neural networks with non-linear units. This method is designed to minimize the total error of the output computed by the network. In such a network, there is an input layer, an output layer, with one or more hidden layers in between them. During training, an input pattern is given to the input layer of the network. Based on the given input pattern, the network will compute the output in the output layer. This network output is then compared with the desired output pattern. The aim of the back-propagation learning rule is to define a method of adjusting the weights of the networks. Then, the network will give the output that matches the desired output pattern given any input pattern in the training set [7].

The training of a network by back-propagation involves three stages: the feed forward of the input training pattern, the calculation and back-propagation of the associated error, the adjustment of the weight and the biases. The main defects of the BPNN can be described as: slow convergence, difficulty in escaping from local minima, easily entrapped in network paralyses, uncertain network structure. In order to overcome the demerits of BPNN some techniques are introduced and it is mentioned below.

Cheng Hua Li and Soon Cheol Park introduced a new method called MRBP. MRBP (Morbidity neuron Rectify Back-Propagation neural network) [5]. This method is used to detect and rectify the morbidity neurons. This reformative BPNN divides the whole learning process into many learning phases. It evaluates the learning mode used in the phase evaluation after every learning phase. This can improve the ability of the neural network, making it more adaptive and robust, so that the network can more easily escape from a local minimum, and be able to train itself more effectively.

Wei Wang and Bo Yu proposed a combined method called MBPNN and LSA. The MBPNN [6] accelerates the training speed of BPNN and improve the categorization accuracy. LSA can overcome the problems caused by using statistically derived conceptual indices instead of individual words. It constructs a conceptual vector space in which each term or document is represented as a vector in the space. It not only greatly reduces the dimension but also discovers the important associative relationship between terms. The two methods to improve the speed of training for BPNN in order to improve the back propagation algorithm in terms of faster convergence and global search capabilities are:

- Introduce momentum into the network. Convergence is sometimes faster if a momentum term is added to the weight update formulas
- Using adaptive learning rate to adjust the learning rate. The role of the adaptive learning rate is to allow each weight to have its own learning rate, and to let the learning rate vary with time as training progress.

Latent semantic analysis (LSA) uses singular value decomposition (SVD) [8] technique to decompose a large term-document matrix into a set of k orthogonal factors, it can transform the original textual data to a smaller semantic space by taking advantage of some of the implicit higher-order structure in associations of words with text objects .These derived indexing dimensions, rather than individual words, can greatly reduce the dimensionality and have the semantic relationship between terms. So even two documents don't have any common words, we also can find the associative relationship between them, because the similar contexts in the documents will have similar vectors in the semantic space. The SVD used for noise reduction to improve the computational efficiency in text categorization and also LSA expanded term by document matrix used in conjunction with background knowledge in text categorization. The supervised LSA had been proposed to improve the performance in text categorization.

MBPNN overcomes the slow training speed problem in the traditional BPNN and can escape from the local minimum. MBPNN enhances the performance of text categorization. The introducing of LSA not only reduces the dimension, further improves its accuracy and efficiency. Bo Yu *et al.* [40] have proposed text categorization models using back-propagation neural network (BPNN) and modified back-propagation (MBPNN). A major problem of text representation is the high dimensionality of feature space. Dimensionality reduction and semantic vector space generation was achieved using a technique Latent Semantic Analysis (LSA). They have tested their categorization models using LSA on newsgroup dataset. They found that computation time for neural network with LSA method was faster than the neural network with VSM model. Further, the categorization performance of neural network using LSA was better than using VSM.

III. Latent Semantic Indexing

Latent Semantic Indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called Singular Value Decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts.

LSI overcomes two of the most severe constraints of Boolean keyword queries: multiple words that have similar meanings (synonymy) and words that have more than one meaning (polysemy). Synonymy and polysemy are often the cause of mismatches in the vocabulary used by the authors of documents and the users of information retrieval systems. [8] As a result, Boolean keyword queries often return irrelevant results and miss information that is relevant.

LSI is also used to perform automated document categorization. In fact, several experiments have demonstrated that there are a number of correlations between the way LSI and humans process and categorize text [9]. Document categorization is the assignment of documents to one or more predefined categories based on their similarity to the conceptual content of the categories [8]. LSI uses example documents to establish the conceptual basis for each category. During categorization processing, the concepts contained in the documents being categorized are compared to the concepts contained in the example items, and a category is assigned to the documents based on the similarities between the concepts they contain and the concepts that are contained in the example documents. Dynamic clustering based on the conceptual content of documents can also be accomplished using LSI. Clustering is a way to group documents based on their conceptual similarity to each other without using example documents to establish the conceptual basis for each cluster. This is very useful when dealing with an unknown collection of unstructured text.

Yan Huang described about Text Categorization via Support Vector Machines (SVMs) approach based on Latent Semantic Indexing (LSI) [10]. Latent Semantic Indexing is a method for selecting informative subspaces of feature spaces with the goal of obtaining a compact representation of document. Support Vector Machines [3] are powerful machine learning systems, which combine remarkable performance with an elegant theoretical framework. The SVMs well fits the Text Categorization task due to the special properties of text itself. The LSI+SVMs frame improves clustering performance by focusing attention of Support Vector Machines onto informative subspaces of the feature spaces. LSI is an effective coding scheme and It captures the underlying content of document in semantic sense. SVMs well fit for text categorization task due to the properties of text. LSI+SVMs shows to be a promising scheme for TC task.

Chung-Hong Lee *et al*. described that an LSI is a technique for Information Retrieval, especially in dealing with polysemy and synonymy [11]. LSI use SVD process to decompose the original term-document matrix into a lower dimension triplet. The triple is the approximation to original matrix and can capture the latent semantic relation between terms. A novel method for multilingual text categorization using Latent Semantic Indexing is mentioned here. The centroid of each class has been calculated in the decomposed SVD space. The similarity threshold of categorization is predefined for each centroid. Test

documents with similarity measurement larger than the threshold will be labeled Positive or else would be labeled Negative. Experimental result indicated that the performance on the precision, recall is quite good using LSI technique to categorize the multi-language text.

Sarah Zelikovitz and Finella Marquez presented a work that evaluates background knowledge created via web searches might be less suitable. For some text classification tasks, unlabeled examples might not be the best form of background knowledge for use in improving accuracy for text classification using Latent Semantic Indexing (LSI) [12]. LSI's singular value decomposition process can be performed on a combination of training data and background knowledge. The closer the background knowledge is to the classification task, the more helpful it will be in terms of creating a reduced space that will be effective in performing classification. Using a variety of data sets, evaluate sets of background knowledge in terms of how close they are to training data, and in terms of how much they improve classification.

Antony Lukas *et al*. made a survey about document categorization using Latent semantic indexing [13]. The purpose of this research is to develop systems that can reliably categorize documents using the Latent Semantic Indexing (LSI) technology [11]. Categorization systems based on the LSI technology do not rely on auxiliary structures and are independent of the native language being categorized .Three factors led us to undertake an assessment of LSI for categorization applications. First, LSI has been shown to provide superior performance to other information retrieval techniques in a number of controlled tests [8]. Second, a number of experiments have demonstrated a remarkable similarity between LSI and the fundamental aspects of the human processing of language. Third, LSI is immune to the nuances of the language being categorized, thereby facilitating the rapid construction of multilingual categorization systems. The emergence of the World Wide Web has led to a tremendous growth in the volume of text documents available to the open source community. It had led to an equally explosive interest in accurate methods to filter, categorize and retrieve information relevant to the end consumer. Of special emphasis in such systems is the need to reduce the burden on the end consumer and minimize the system administration of the system. The implementation of two successfully deployed systems employing the LSI technology for information filtering and document categorization was described. The systems utilize in-house developed tools for constructing and publishing LSI categorization spaces.

Two-stage feature selection algorithm [32] is based on a kind of feature selection method and latent semantic indexing. Feature selection is carried out in two main steps. First, a new reductive feature space is constructed by a traditional feature selection method. In the first stage, the original features dimension is decreased from m to t. Second, features are selected by LSI method on the basis of the new reductive feature space that was constructed in the first stage. In the second stage, the features dimension is decreased from t to k. The feature-based method and semantic method are combined to reduce the vector space. The algorithm not only reduces the number of dimensions drastically, but also overcomes the problems existing in the vector space model used for text representation.

I.Kuralenok and I. Nekrest'yanov [41] have considered the problem of classifying the set of documents into given topics. They have proposed a classification method based on the use of LSA to reveal semantic dependencies between terms. The method used the revealed relationships to specify the function of the topical proximity of terms, which was then used to estimate the topical proximity of documents. The results indicated a high quality of classification. The computation cost of this method was high at the initial stage and relatively cheap at the classification stage. However, considering the problem of clusterization of documents, unlike the classification problem, the topics of the groups are not given in advance.

IV. Support vector machines

SVMs are a set of related supervised learning methods used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. A support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class since in general the larger the margin the lower the generalization error of the classifier [14].

Lukui Shi *et al*. proposed an algorithm combined nonlinear dimensionality reduction techniques with support vector machines for text classification. To classify documents, the similarity between two text documents is considered in many algorithms of text categorization. Here geodesic distance is used to represent the similarity between two documents. In this algorithm, high-dimensional text data are mapped into a low-dimensional space with the ISOMAP algorithm after geodesic distances among all documents are computed at first. Then the low-dimensional data are classified with a multi-class classifier based single-class SVM [15].

ISOMAP is a nonlinear dimensionality reduction technique, which generalizes MDS by replacing Euclidean distances with an approximation of the geodesic distances on the manifold. The algorithm is to compute the geodesic distances between points, which represent the shortest paths along the curved surface of the manifold. For neighboring points; the input space distance gives a good approximation to the geodesic distance. For objects, the geodesic distances can be approximated by a sequence of short hops between neighboring points.

The multi-class classifier based on single-class SVM can effectively treat multi-class classification problems. The efficiency of the classifier will be rapidly degraded when the dimension of data becomes greatly high. Usually, the dimension of text data is huge. To fast classify high-dimensional text data, it is necessary to decrease the

dimension of high-dimensional data before classifying text documents. It is a good selection to combine the above multi-class classifier with ISOMAP.

Montanes E *et.al.* described a wrapper approach with support vector machines for text categorization [16]. Text Categorization is the assignment of predefined categories to documents plays an important role in a wide variety of information organization and management tasks of Information Retrieval (IR). It involves the management of a lot of information, but some of them could be noisy or irrelevant and hence, a previous feature reduction could improve the performance of the classification. Here they proposed a wrapper approach. This approach is time-consuming and also infeasible. But this wrapper explores a reduced number of feature subsets and also it uses Support Vector Machines (SVM) [18] as the evaluation system; and these two properties make the wrapper fast enough to deal with large number of features present in text domains.

István Pilászy [17] gave a short introduction of text categorization (TC), and important tasks of a text categorization system. He also focused on Support Vector Machines (SVMs), the most popular machine learning algorithm used for TC.

Support Vector Machines (SVMs) have been proven as one of the most powerful learning algorithms for text categorization. Support vector machines (SVMs) [19] are a set of related supervised learning methods used for classification and regression. In simple words, given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [20].Redundant features and high dimension are well-handled.

Linear Support Vector Machines (SVMs) [33] have been used successfully to classify text documents into set of concepts. The training time was taken with respect to each category by SVM*light*, PSVM, SVMlin, and SVMperf on two corpuses. The training times of all other algorithms were higher than SVM *light* on both corpuses. On reuters-21578, the training time of PSVM is the least, and on assumed, both SVMlin and PSVM achieve less training time when compared with other algorithms. The order of computational complexity of PSVM scales with respect to dimensionality of the corpus. The solution of FPSVM can also be obtained by solving system of simultaneous linear equations similar to PSVM. PSVM maintains almost constant training time irrespective of the penalty parameter and categories. The performance of PSVM can greatly be improved by using it along with advanced feature selection/extraction methods like word clustering, rough sets.

Wenqian Shangahan *et al.* [38] have designed a novel Gini index algorithm to reduce the high dimensionality of the feature space. They have constructed a new measure function of Gini index to fit text categorization. Improved Gini index algorithm was evaluated using three classifiers: SVM, kNN, fkNN. The performance of new Gini index was compared with feature selection methods Inf Gain, CrossEntroy, CHI, and Weigh of Evid. The results showed that the performance of new method was best in some dataset and inferior in another dataset. As a whole, they concluded that their improved Gini index showed better categorization performance.

V. Decision Tree

Decision tree learning [25], used in data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. A decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions rather the resulting classification tree can be an input for decision making [24].

The text categorization performance of purely inductive method is used [23].Two inductive learning algorithms are: Bayesian classifier and other one is Decision tree. Both the algorithms studied about indexing the data for document retrieval and also extraction of data from the text sources.

The Bayes rule is to estimate the category assignment probabilities and then assign to a document those categories with high probabilities. The decision tree use the algorithm DT-min 10: to recursively subdivide the training examples into subsets based on the information gain metric [21].

Maria Zamfir Bleyberg and Arulkumar Elumalai introduced a rough set method. It is founded on the assumption that with every object of the universe we associate some information. Objects characterized by the same information are similar in view of the available information about them. The indiscemibility relation generated in this way is the mathematical basis of rough set theory. Any set of all indiscernible objects is called an elementary set, and forms a basic granule of knowledge about universe. Any union of some elementary sets is referred as crisp set, otherwise the set is rough. In the rough set theory, any vague concept is replaced by a pair of precise concepts: the lower and the upper approximation of the vague concept. The learning methods based on rough sets, can be used to support flexible, dynamic, and personalized information access and management in a wide variety of tasks.

VI. Naive Bayes Classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. A naive Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. One can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified

assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. An advantage of the naive Bayes classifier is that it requires a 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 [26].

Jantima Polpinij and Aditya Ghose solved an ambiguity problem of software errors because much of the requirements specification is written in a natural language format. It is hard to identify consistencies because this format is too ambiguous for specification purposes. [27] A method for handling requirement specification documents which have a similar content to each other through a hierarchical text classification. The method consists of two main processes of classification: heavy classification and light classification. The heavy classification is to classify based on probabilistic text classification (Naïve Bayes), while light classification is to handle elaborate specification requirement documents by using the Euclidean Distance. Slimming down the number of requirements specification through hierarchical text classification classifying may yield a specification which is easier to understand. That means this method is more effective for reducing and handling in the requirements specification.

Dino Isa *et al.* [42] have designed and evaluated a hybrid classification approach by integrating the naïve Bayes classification and SOM utilizing the simplicity of the naïve Bayes to vectorize raw text data based on probability values and the SOM to automatically cluster based on the previously vectorized data. Through the implementation of an enhanced naïve Bayes classification method at the front-end for raw text data vectorization, in conjunction with a SOM at the back-end to determine the right cluster for the input documents, better generalization, lower training and classification time, and good classification accuracy was obtained. The drawback of this technique is the fact that the classifier will pick the highest probability category as the one to which the document is annotated too.

VII. Self-Organizing Map

The SOM [28] is an unsupervised-learning neural-network method that produces a similarity graph of input data. It consists of a finite set of models that approximate the open set of input data, and the models are associated with nodes (neurons) that are arranged as a regular, usually 2-D grid. The models are produced by a learning process that automatically orders them on the 2-D grid along with their mutual similarity.

Cheng Hua Li and Soon Choel Park described two kinds of neural networks for text categorization [30], multi-output perceptron learning (MOPL) and back-propagation neural network (BPNN), and then a novel algorithm using improved back-propagation neural network is proposed. This algorithm can overcome some shortcomings in traditional back-propagation neural network such as slow training speed and easy to enter into local minimum. The training time and the performance, and tested three methods are compared. The results showed that the proposed algorithm is able to achieve high categorization effectiveness as measured by the precision, recall and F-measure.

Richard Freeman *et al.* [35] have investigated the use of self-organizing maps for document clustering. They have presented a hierarchical and growing method using a series of one-dimensional maps. The documents were represented using vector-space model. Dynamically growing one-dimensional SOM were allocated hierarchically to organize the give set of documents. The hierarchical structured maps produced were visualized easily as a hierarchical tree. The results showed a more intuitive representation of a set of clustered documents.

Nikolaos and Stavros [36] have introduced LSISOM method, for automatic categorization of document collections. The method LSISOM obtained word category histograms from the SOM clustering of the Latent Semantic Indexing representation of document terms. The problem of high dimensionality of VSM word histograms document representation was suppressed by LSI representation. The SOM used was a two-dimensional SOM. They used 420 articles as dataset from the TIME Magazine. They have proved that LSISOM method is computationally efficient due to dimensionality reduction using LSI of documents. They have compared Standard SOM (SSOM) and LSISOM for a set of documents. They justified that consistent mapping of documents onto a single cluster was obtained by LSISOM.

The method topological organization of content (TOC) [37] is topology preservation of neural network for content management and knowledge discovery. TOC generate taxonomy of topics from a set of unstructured documents. TOC is a set of 1D-growing SOMs. The TOC method produced a useful hierarchy of topics that is automatically labeled and validated at each level. This approach used entropy–based BIC to determine optimum number of nodes. TOC and 2D-SOM were compared; the results showed that topological tree structure improved navigation and visualization. The main advantages of the approach are the validation measure, scalability, and topology representation. To improve TOC, feature selection method LSA can be used to enhance the association between terms.

Yan Yu et *al.* [39] have presented a new document clustering method based on one-dimensional SOM. This method obtained the clustering results by calculating the distances between every two adjacent MSPs (the most similar prototype to the input vector) of well trained 1D-SOM. Their work proved that procedure using 1D-SOM is simple and easy relative to that with 2D-SOM.

Tommy W. S. Chow and M. K. M. Rahman [43] have proposed a new document retrieval (DR) and plagiarism detection (PD) system using multilayer self-organizing map (MLSOM). Instead of relying on keywords/lines, the proposed scheme compared a full document as a query for performing retrieval and PD. The tree-structured representation hierarchically includes document features as document, pages, and paragraphs. MLSOM, a kind of extended SOM model, was developed for processing tree-structured data. A tree data consists of nodes at different levels. In MLSOM, there were as many SOM layers as the number of levels in the tree. They mapped the position vectors of child nodes into the SOM input vector. The mapping of position vectors was conducted using a simple 1D- SOM that is trained. Experimental results using MLSOM were compared against

tree-structured feature and flat-feature. They have shown that tree-structured representation enhanced the retrieval accuracy and MLSOM served as an efficient computational solution. However, for a very large scale implementation of DR and PD, it is difficult to process all documents in a single MLSOM module.

VIII. Genetic Algorithm

Genetic Algorithm is a search technique based on the principles of biological evolution, natural selection, and genetic recombination. They simulate the principle of 'survival of the fittest' in a population of potential solutions known as chromosomes. Each chromosome represents one possible solution to the problem or a rule in a classification.

The population evolves over time through a process of competition whereby the fitness of each chromosome is evaluated using a fitness function. During each generation, a new population of chromosomes is formed in two steps. First, the chromosomes in the current population are selected to reproduce on the basis of their relative fitness. Second, the selected chromosomes are recombined using idealized genetic operators, namely crossover and mutation, to form a new set of chromosomes that are to be evaluated as the new solution of the problem. GAs are conceptually simple but computationally powerful. They are used to solve a wide variety of problems, particularly in the areas of optimization and machine learning [29].

Clustering is an efficient way of reaching information from raw data and K-means is a basic method for it. Although it is easy to implement and understand, K-means has serious drawbacks. Hongwei Yang had presented an efficient method of combining the restricted filtering algorithm and the greedy global algorithm and used it as a means of improving user interaction with search outputs in information retrieval systems [31]. The experimental results suggested that the algorithm performs very well for Document clustering in web search engine system and can get better results for some practical programs than the ranked lists and k-means algorithm.

Wei Zhao *et.al.* introduced a new feature selection algorithm in text categorization [34]. Feature selection is an important step in text classification, which selects effective feature from the feature set in order to achieve the purpose of reduce feature space dimension. Genetic algorithm (GA) optimization features are used to implement global searching, and k-means algorithm to selection operation to control the scope of the search, which ensures the validity of each gene and the speed of convergence. Experimental results show that the combination of GA and k-means algorithm reduced the high feature dimension, and improved accuracy and efficiency for text classification.

IX. Conclusion

This paper discusses about various classification algorithms, their merits and demerits. The data categorization includes the categorization of text, image, object, voice etc. The focus of survey is done mainly on text categorization. The representation techniques, supervised and unsupervised classification algorithms and their applications are discussed. The survey has shown that different techniques exist for the problem. The research should be still concentrated on efficient feature selection and on categorizing different types of data in different fields. In order to improve the text categorization various other semantic based machine learning algorithms can be added in future.

References

[1] M.-L. Antonie and O. R. Za¨ıane, "Text document categorization by term association", In Proc. of the IEEE 2002 International Conference on Data Mining", pp.19–26, Maebashi City, Japan, 2002.

[2] Yiming Yang and Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", CiteSeerX, 1997.

[3] Thorsten Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features, Machine Learning: ECML-98, Vol.1398, pp.137-142, 1998.

[4] Nidhi and Vishal Gupta, "Recent Trends in Text Classification Techniques", International Journal of Computer Applications, Vol.35, No.6, 2011.

[5] Cheng Hua Li and Soon Cheol Park, "A Novel Algorithm for Text Categorization Using Improved Back-Propagation Neural Network", Springer, pp. 452 – 460, 2006.

[6] Wei Wang and Bo Yu, "Text categorization based on combination of modified back propagation neural network and latent semantic analysis", *Neural Comput & Applic*, Springer Link, Vol. 18, No.8, pp.875–881, 2009.

[7] Wei Wu, Guorui Feng, Zhengxue Li, and Yuesheng Xu, "Deterministic Convergence of an Online Gradient Method for BP Neural Networks", IEEE Transactions on Neural Networks, Vol.16, NO.3, 2005.

[8] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki and Santosh Vempala, "Latent Semantic Indexing: A Probabilistic Analysis", Journal of Computer and System Sciences, Vol.61, pp.217_235, 2000.

[9] S.T. Dumais, G. W. Furnas, T. K. Landauer, and S. Deerwester, "Using latent semantic analysis to improve information retrieval", Proceedings of CHI'88: Conference on Human Factors in Computing, ACM, pp.281_285, 1988.

[10] Yan Huang, "Support Vector Machines for Text Categorization Based on Latent Semantic Indexing", Electrical and Computer Engineering Department, isn.ucsd.edu, 2003.

[11] Chung-Hong Lee, Hsin-Chang Yang and Sheng-Min Ma, "A Novel Multilingual Text Categorization System using Latent Semantic Indexing", Proceedings of the First International Conference on Innovative Computing, Information and Control (ICICIC'06), 2006.

[12] Sarah Zelikovitz and Finella Marquez, "Evaluation of Background Knowledge for Latent Semantic Indexing Classification", American Association for Artificial Intelligence, 2005.

[13] Anthony Zukas and Robert J. Price, "Document Categorization Using Latent Semantic Indexing", Symposium on Document Image Understanding Technologies, 2003.

[14] Daniela Giorgetti and Fabrizio Sebastiani, "Multiclass Text Categorization for Automated Survey Coding", SAC 2003.

[15] Lukui Shi, Jun Zhang, Enhai Liu, and Pilian He, "Text Classification Based on Nonlinear Dimensionality Reduction Techniques and Support Vector Machines", Third International Conference on Natural Computation, IEEE Xplore, Vol.1, pp.674-677, 2007.

[16] Montanes E., Quevedo J. R. and Diaz I., "A Wrapper Approach with Support Vector Machines for Text Categorization", Springer, LNCS 2686, pp. 230-237, 2003.

[17] István Pilászy, "Text Categorization and Support Vector Machines", Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence, 2005.

[18] Manabu Sassano, "Using Virtual Examples for Text Classification with Support Vector Machines", Journal of Natural Language Processing, Vol.13, No.3, pp. 21-35. 2006.

[19] A. Basu, C. Watters and M. Shepherd, "Support Vector Machines for Text Categorization", Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03), Vol.4, pp.103.3, 2003.

[20] Edda Leopold and Jorg Kindermann,"Text Categorization with Support Vector Machines: How to Represent Texts in Input Space?", Machine Learning, Vol.46, Nr.1-3, pp.423–444, 2002.

[21] Chidanand Apt, Fred Damerau and Sholom M. Weiss, "Automated Learning of Decision Rules for Text Categorization", ACM Transactions on Information Systems, 1994.

[22] Srinivasan Ramaswamy, "Multiclass Text classification A Decision Tree based SVM Approach", CS294 Practical Machine Learning Project, Citeseer, 2006.

[23] David D.Lewis and Mark Ringuette, "A comparison of two learning algorithms for text Categorization, Symposium on Document Analysis and Information Retrieval", 1994.

[24] C. Apte, F. Damerau, and S.M. Weiss, "Text Mining with Decision Trees and Decision Rules", Conference on Automated Learning and Discovery Carnegie-Mellon University, 1998.

[25] Nerijus Remeikis, Ignas Skucas and Vida Melninkaite, "A Combined Neural Network and Decision Tree Approach for Text Categorization", Information Systems Development, Springer, pp.173-184, 2005.

[26] P.Bhargavi and Dr.S.Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", IJCSNS International Journal of Computer Science and Network Security, VOL.9, No.8, 2009.

[27] Mohamed Aly, "Survey on Multiclass Classification Methods", Neural Networks, 2005.

[28] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Jukka Honkela, Vesa Paatero, and Antti Saarela, "Self Organization of a Massive Document Collection", IEEE Transactions on Neural Networks, Vol.11, No. 3, pp.574-585, 2000.

[29] Xiaoyue Wang, Zhen Hua and Rujiang Bai, "A Hybrid Text Classification model based on Rough Sets and Genetic Algorithms", SNPD '08. Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE Xplore, pp.971-977, 2008.

[30] Cheng Hua Li and Soon Choel Park, "Text Categorization Based on Artificial Neural Networks", Neural Information Processing, Springer, Vol.4234, pp.302 – 311, 2006.

[31] Hongwei Yang, "A Document Clustering Algorithm for Web Search Engine Retrieval System", International Conference on e-Education, e-Business, e-Management and e-Learning, IEEE, pp.383-386, 2010.

[32] Jiana Meng and Hongfei Lin, "A Two-stage Feature Selection Method for Text Categorization", Seventh International Conference on Fuzzy Systems and Knowledge Discovery, IEEE, pp.1492-1496, 2010.

[33] M. Arun Kumar and M. Gopal, "An Investigation on Linear SVM and its Variants for Text Categorization", Second International Conference on Machine Learning and Computing, IEEE, pp.27-31, 2010.

[34] Wei Zhao, Yafei Wang and Dan Li, "A New Feature Selection Algorithm in Text Categorization", International Symposium on Computer, Communication, Control and Automation, IEEE, pp.146-149, 2010.

[35] Richard Freeman, Hujun Yin and Nigel M. Allinson, "Self-Organizing Maps for Tree View Based Hierarchical Document Clustering", IEEEXplore, pp.1906-1911, 2002.

[36] Nikolas Ampazis and Stavros J. Perantonis, "LSISOM - A Latent Semantic Indexing Approach to Self-Organizing Maps of Document Collections", Neural Processing Letters 19, pp. 157-173, 2004.

[37] Freeman R.T. and Hujun Yin, "Web Content Management by Self Organization", IEEE transactions on Neural Networks, Vol.16, No.5, pp.1256-1268, 2005.

[38] Wenqian Shang, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu and Zhihai Wang, "A novel feature selection algorithm for text categorization", Expert Systems with Applications 33, pp.1-5, 2007.

[39] Yan Yu, Pilian He, Yushan Bai and Zhenlei Yang, "A Document Clustering Method Based on One-Dimensional SOM", Seventh IEEE/ACIS International Conference on Computer and Information Science, pp.295-300, 2008.

[40] BoYu, Zong-ben Xu and Cheng-hua Li, "Latent semantic analysis for text categorization using neural network", Knowledge-Based Systems, Vol.21, pp.900-904, 2008.

[41] Kuralenok I. and Nekrest'yanov I., "Automatic Document Classification Based on Latent Semantic Analysis", Programming and Computer Software, Vo.26, No.4, pp.199-206, 2000.

[42] Dino Isa, Kallimani V.P. and Lam Hong Lee, "Using the self organizing map for clustering of text documents", Expert Systems with Applications, Vol.36, pp.9584-9591, 2009.

[43] Tommy W.S. Chow and Rahman M.K.M., "Multilayer SOM with Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection", IEEE Transactions On Neural Networks, Vol.20, No.9, pp.1385-1402,2009.