Journal of Theoretical and Applied Information Technology
30th September 2016. Vol.91. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TABLE-BASED MATCHING APPROACH USING GENETIC ALGORITHM FOR FEATURE SELECTION IN TEXT CATEGORIZATION

[1]B. SUNIL SRINIVAS, [2]A. GOVARDHAN

[1]Assistant Professor, CSE Dept., VVIT, Hyderabad, & Research Scholar, CSE Dept., JNTUA
[2]Professor, CSE Dept., JNTU College of Engineering, JNTUH, Hyderabad
Email: [1]sunilsrinivas16@gmail.com, [2]govardhan_cse@yahoo.co.in

ABSTRACT

Text categorization is a significant approach to manage the increasing text data on the Internet and is a significant research issue since the out bursting of digital and online web where numerous documents are available online and has been increased greatly in recent years. In this paper, a modified version of Table aided Matching algorithm for text categorization is proposed. This approach addressed the issue of huge dimension in order to maximize the computational efficiency and accuracy. The genetic algorithm has the ability to solve this approach. Thus, prior to classification, the dimensionality reduction technique is employed where the size of the documents in each profile is minimized. The performance evaluation of the suggested approach is matched with the two existing classification methodologies and has been demonstrated that the proposed approach has better results matched with existing approaches.

Keywords: *Text Categorization, Feature Selection, Genetic Algorithm, Table-based Matching Approach, Classification.*

1. INTRODUCTION

Text Categorization is becoming a significant research problem since the increasing of digital and online web data, where numerous documents are available online and has been increased greatly in recent years. Text Categorization is defined as the procedure of spontaneously allotting a tag to the specified uncategorized text. Memory is becoming inexpensive by accumulating the storage ability and transmission rate speed. CPU processing capability is likewise becoming further dominant. This technological enhancement provide more information that is essential to be processed and obtains relevant information in time that is important for many applications where text categorization is the key issue [2].

Text categorization has been used in many applications such as spam email filtering [3], Web page classification [4], customer relationship management [5], and text sentinel classification [6]. In the current eras, numerous text categorization approaches are being developed to allocate text documents to their annotated categories, such as Bayesian [8], support vector machine, artificial neural networks [9], Rocchio [10], latent Dirichlet allocation [11], and many other machine learning and statistical approaches [12].

Feature selection is an essential phase in text categorization where the relevant features are selected from the feature set so as to attain the reduced feature space magnitude ,that is to pick a number d such that (d<D) set of optimum feature subset in D feature set. The conventional feature selection approaches employed are Document Frequency, Information Gain, and Mutual Information etc.

1.1 Motivation

In the Traditional Approach, Text Categorization is initialized by encoding the documents into arithmetical vector and causes two key problems: first is huge dimensionality and the other is the sparse distribution of words on the documents. This issue leads to expensive cost for executing every arithmetical vector signifying a document in terms of time and system resources. Over 90% of features in vector space are zero values in every arithmetical vector that degraded the discrimination amongst arithmetical vectors and causes poor performance.

This has been overcome by constructing the tables for each documents instead of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

numerical vectors. The table comprises of the word in the document and its associated weight corresponding to all the documents. Though sparse distribution is minimized, huge dimensionality still exists in this approach where the computational efficiency & accuracy is minimized. The higher magnitude not merely augment the executing time of building classifiers but also minimizes the accuracy of classification outcomes.

So as to resolve the issues of higher dimension, diminishing attributes is very essential and critical for enhancing the efficacy and efficiency of text categorization. Therefore, the proposed paper addressed this issues in order to minimize the huge dimension and maximize the computational efficiency and accuracy. Genetic Algorithm [13] has the capability to resolve the issue of resembling optimum solution in space, that employs the genetic operators to explore the text content, uses fitness value to estimate the advantages of the candidate outcomes. By means of which the categorization of the new unseen document are done using selected features of the documents. Thus, the computational efficiency and accuracy is maximized compared to the existing Normalized Table Matching algorithm for text categorization [1].

1.2 Organization of the paper

A brief introduction to the text categorization, features selection along with the motivation for the approach is given in this section. Section 2 briefly discusses the existing issues in the text categorization using different machine learning approaches. The genetic algorithm is briefly described in the section 3. The proposed table-based matching approach using genetic algorithm is briefly discussed in section 4. The experimental outcomes and its brief investigation for the test bed is given in section 5 followed by conclusions and references given in section 6 and section 7 correspondingly.

2. LITERATURE SURVEY

In [16] an efficient self-constructing fuzzy feature clustering (SCFFC) approach is proposed for text categorization. This methodology introduced three diverse approaches, hard-weighting, soft-weighting, and mixed-weighting, to obtain attributes for learning text categorization approaches. The methodology clusters words into groups depending on the dissemination of class labels accompanied with every word. The survey on clustering the words could be obtained in [7]. The other form of methodology suggested in [15], specifies a feature selection approach by means of correlation coefficient clustering. This approach gathers the attributes into groups by evaluating its correlation coefficients, and formerly the utmost class-dependent attribute in every group is nominated. Further feature selection approaches could be obtained in [14].

The incremental orthogonal centroid (IOC) approach is the other feature extraction approach suggested in [19]. This approach presumes that a document set D is an n |D| matrix. The matrix formerly attempts to obtain an optimum transformation matrix to transform the original matrix into an n k matrix, where k << |D|. The discriminant constant is a feature extraction approach suggested in [17, 18], which inherits the benefits of matrix transformation excluding the transformed original matrix into a k |D| matrix. In [20], a kernel function of two raw texts is suggested for employing SVM for text categorization [20]. An outcome of its suggested kernel function specifies a syntactic resemblance amongst two raw texts as two elongated strings. In [21], two alternates of the EM algorithm is suggested for soft clustering, where each element is permissible to fit to more than single group, and employed them to text clustering and gene expression grouping [21].

In [22], an alternate method to machine learning approaches for classifying news articles defined as simple texts. The two issues are evaded by encrypting a text or texts into a table, as an alternative to arithmetical vectors. The similarity matrix dependent form of NTSO as the methodology to the text classification is given in [23]. This is suggested to encrypt texts into string vectors and employ the NTSO (Neural Text Self Organization) as the string vector aided neural network for the text categorization. By encrypting texts into the other form, it is attempted to evade the two significant issues, entirely.

In [24], an improved form of single pass approach specified for text categorization is suggested. In this form, documents are mapped into tables and the operation on two tables are specified for employing the single pass approach. The objective of this methodology is to enhance the performance of single pass approaches for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

text categorization by altering it into the specific form. In [1] documents are encrypted into normalized tables for classifying data spontaneously. Priori, the table aided methodology was recommended, the categorical values represents how much the document is appropriate to the specified category that might be overestimated or underestimated by the given document dimension. As an outcome to this issue, document is encrypted into stable sized tables, specifying the operation for calculating the similarity amongst two tables like a normalized value, and categorized it arithmetically. As an advantage from this study, it is permitted to evaluate the category scores autonomously of a specified document dimension, considering the weights from both documents, and assume the further constant performance.

3. GENETIC ALGORITHM

Genetic algorithms (GAs) are well known for its capability to competently explore huge domain about which little is known earlier. As GAs are comparatively unresponsive to noise, they appear to be an outstanding selection for the origin of a further strong feature selection approach for augmenting the efficiency of the texture classification. GAs is a kind of inductive learning approach and is an adaptive exploration method that revealed significant enhancement over a diversity of arbitrary and local exploring procedures. This is accomplished by its capability to exploit gathered data regarding a primarily unidentified exploration domain so as to bias succeeding investigation into encouraging subdomains. As GAs are essentially a domain free searching procedure, they are perfect for applications where domain understanding and idea is problematic or unbearable to offer.

The foremost issues in employing GA to any problem is choosing a suitable illustration and an acceptable evaluation function. GA is a stochastic approach that simulates natural evolution. The utmost discrete feature of this approach is that it conserves a group of outcomes known as individuals or chromosomes in a populace. Similar to natural evolution, it is a procedure of choosing suitable chromosomes at every iteration. To execute the procedure of evolution, the nominated chromosomes go through genetic functions, like crossover and mutation. Additionally, different from numerous exploration approaches, that executes a local, greedy exploration, GAs performs a global exploration. The GAs mimics the procedures in biological systems for evolutions depending on the notion of "survival of the fittest" defined by Charles Darwin.

In GA, reproduction picks worthy set of input features, crossover amalgamates good strings as to produce enhanced individuals and mutation modifies a string closely to endeavor as to generate an enhanced individuals. In every iteration, the population is estimated and verified for termination of the methodology. If the stopping strategy is not achieved, the populace is employed by means of three GA operators and then is re-estimated. This methodology is repeated for given number of iterations. The traditional Genetic Approach is given in figure 1 below.

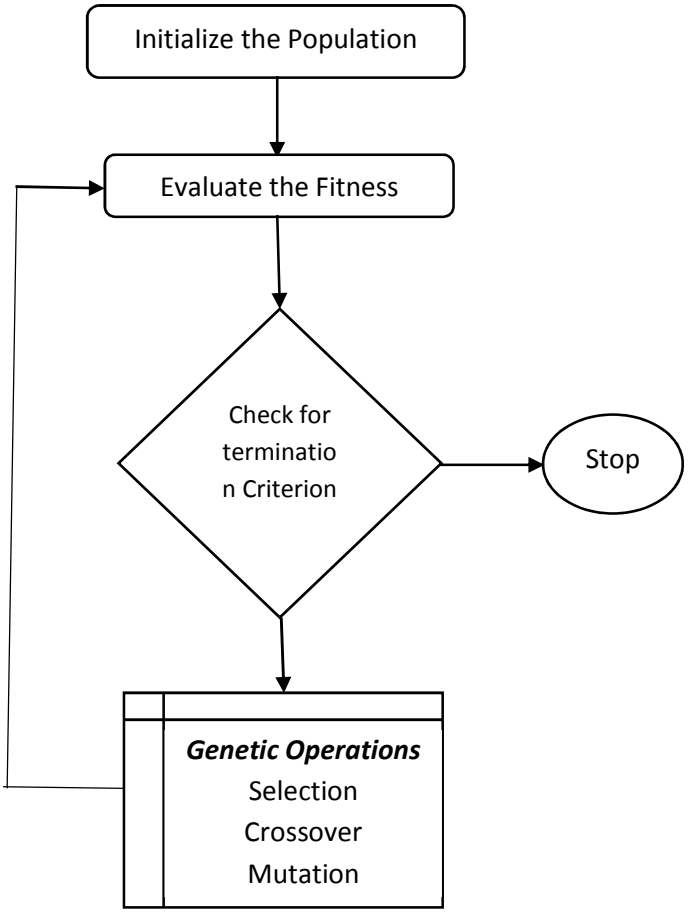

Fig 1: Flow Chart Of Genetic Algorithm

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. TABLE-BASED MATCHING USING GA APPROACH

In this paper, the issue of huge dimensionality which exists in the Normalized Table-based Approach [1] is addressed, where this huge dimensionality is reduced by feature selection approach using heuristic approach like genetic algorithm. Huge Dimensionality means, every document has higher magnitude mathematical vector for averting the data loss and also consumes more time for implementation. Numerous training samples are essential for the healthy classification relative to the magnitude. Therefore, in existing Normalized Table based Matching approach [1], the text is encoded into tables, where every word is represented using its weight values. The proposed approach functions in the following phases:

A. Encoding Phase

The encoding of the document is done through five steps. They are:

1. *Tokenization*: The complete document is segmented into tokens with the help of space or punctuation marks that produces a group of tokens. The token refers to a word in its raw form.
2. *Stemming & Exceptional Handling*: Each token of the document is converted to its original form by reducing it or employing an exceptional regulation to it.
3. *Removal Stop Words*: The stop words are the grammatical words that has merely grammatical operations and is irrelevant to the data of the text document. Stop words are removed to process the document more efficiently.
4. *Removal of Redundant Words*: The recurrent words in the table are removed as to minimize the redundancy.
5. *Weight Computation*: The weight of every word is computed and listed in the table. Weights of the word specifies how much significant it is in terms of the relevancy to the data of the document. The weight is calculated as follows:

$$Weigth_i(w_k) = tf_i(w_k)(log_2 D - log_2 df(w_k) + 1$$

Here $Weigth_i(w_k)$ is the weight of the term w_k relevantly to the document *i*, $tf_i(w_k)$ is the number of occurrences of the word w_k, D is the total number of documents in the database including word w_k

B. Learning Phase

This phase is technically referred to as the learning phase which is later employed for classification. This is the procedure of constructing guidelines or equations of grouping by means of sampled categorized documents in perspective of text categorization. Learning is essential for categorization. In this phase, sample document is allocated corresponding to the categorization, according to the labels given in the text corpus. From the collection of documents labelled identically, table is constructed and known as categorical profile each with N number of documents in each profile. The total number of categorical profiles are equal to the number of diverse classes that are present in the corpus. Learning is performed by assigning the concatenation which combines complete text of documents into single text document.

C. Dimensionality Reduction Phase

This is novel phase that is introduced in between the learning and classification phase for minimization of documents size in each categorical profile. Apart from the filtering step that is present in [1], this paper introduced Genetic Algorithm for the reduction of the constructed table size and then applied for text classification. Even though the Filtering technique minimizes the table dimension depending on the weights considering that the information loss is minimum, it always does not support to obtain the minimized table with less information loss. Thus, genetic algorithm for dimensionality reduction is applied at this step. The working of traditional genetic algorithm is given in section 3.

1. The first step in genetic approach is initialization of population size. The documents with the table format are taken as the initial population in this approach and the words within the table are chromosomes of the approach.
2. The second step is the calculation of the objective function for minimizing the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

table size. The objective or fitness function is the maximization of the similarity in between the two documents. The evaluation of similarity function is given as:

$$Sim(D_A, D_B) = \frac{W_{CA} + W_{CB}}{W_A + W_B}$$

Here, W_{CA} and W_{CB} indicates the sum of the weights of the common words from document table A and document table B respectively. W_A and W_B indicates the sum of the weights of the words of the document tables A and B respectively.

3. The third step is applying genetic operations on the documents with minimal similarity. The three genetic operations are: Selection, Crossover and Mutation
   - The selection between the documents amongst the corpus is done randomly. Any two documents are selected in order to perform the crossover and mutation operations.
   - In this approach, the crossover operation is performed by interchanging the specified set of words with in two different documents.
   - The mutation operation is performed by inserting the new word into the existing two documents.
4. The fourth step is the termination criterion. The termination criterion in this approach is the maximum number of generations. The complete algorithm terminates until the maximum number of generations are reaches.

Finally, with this approach, the documents with minimum similarity function is removed and documents with maximum similarity function are considered for further classification. This, the dimension of each categorical profile is reduced depending on the relevancy without information loss.

D. Classification Phase

Classification denotes to the procedure of categorizing an undetected document depending on the specified instructions or equations. In this phase, classification is the approach of determining one of the predetermined groups to every undetected document. The undetected document is also encrypted into table and the similarities of the table with the categorical outlines determined as tables are estimated. Thus, the undetected document is classified into the category corresponding to the maximal similarity among its tables and equivalent categorical profile.

5. EXPERIMENTAL RESULTS AND ITS ANALYSIS

This section is deals with the experimentation for assessing the efficiency of the suggested methodology to the jobs of text classification. The evaluation metric that is employed in this approach is the F1 measure which is estimated as given below:

$$F1 = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

The precision is the proportion of truthfully categorized instances with the positive class to the number of elements categorized with the positive class by the defined classifier, and the recall is the proportion of truthfully categorized instances to the number of the positive elements.

5.1 Dataset Description

The experiments is carried out on 20NewsArticles dataset. This dataset has higher categories of news articles. Data are preset into 100-dimensional arithmetical vectors and with the dimension of size 10 in tables. In subsequent section, the collection of news articles known as 20NewsGroups are described, the empirical outcomes are presented and discussed.

The group of news articles known as '20NewsArticle' are employed as the dataset group for estimating the methodologies obtained from, http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html. In the collection, completely 20,000 news articles and 20 clusters are specified. The 20 groups are specified as the two-level hierarchical format where the initial stage has the four groups and the subsequent stage has the 20 groups. In this group of experimentation, four groups are picked such as computer, record, natural science, and social science. In every class, the training group comprises of 2,800 news articles in the sequential order with 700 articles per each class

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and the test set comprises of 1,200 articles, 300 articles per each category.

5.2 Results and its Analysis

The Genetic Algorithm is executed on each categorical profile documents separately. The parameter for the execution of genetic algorithm in specified in the table 1.

Table 1: Parameters For Genetic Algorithm

Parameters	Value
Crossover probability	65%
Crossover point	Arbitrarily Nominated Single-point Crossover
Mutation probability	0.50%
Population Size	50 documents
Selection	Roulette-Wheel Selection
Termination Criterion	100 number of generations

The proposed approach is compared with the two existing classification algorithms. They are Support Vector Machine Classification Approach and Normalized Table-Matching Algorithm for text categorization. 75% of the documents in each categorical profiles 1 to M as the best featured documents using Genetic Algorithm is selected in this approach for Classification. The F1 Measure Averages and Variances for all the three classification approaches are given in the table 2.

Table 2: Comparison Of Performance Analysis Of The Proposed Approach

	SVM	Normalized Table-based Algorithm	Proposed Table-Based Approach
F1 Average	0.6621	0.7801	0.8245
F1 Variance	0.0001520	0.008028	0.007656

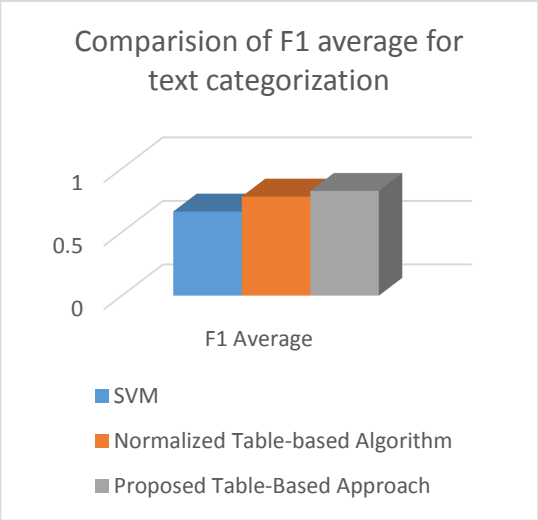

Fig 3: Comparison Of F1 Average For Text Categorization

The overall implementation of the three methodologies traversing over the four groups are suggested in Figure 3 and Figure 4. As shown Table 2, the suggested methodology has the highest F1 measure compared to the other two methods. Distinct from preceding group of experiments, the approach has greater variance and its steadiness is minimum compared to the two approaches.

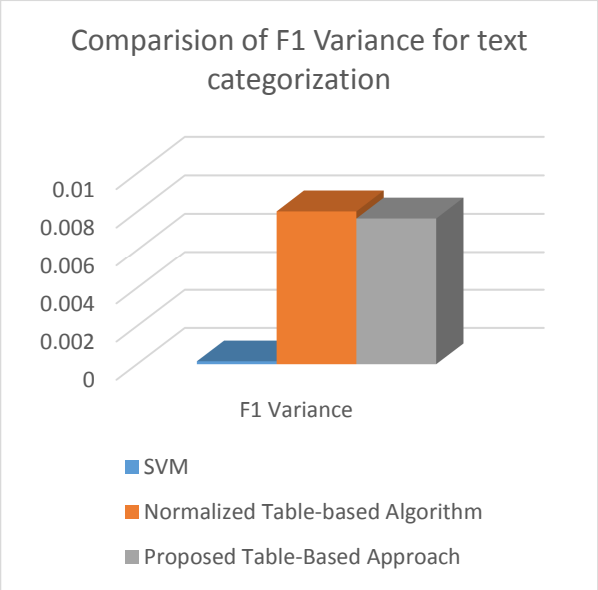

Fig 4: Comparison Of F1 Variance For Text Categorization

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. CONCLUSIONS

This paper suggested a substitution method to the machine learning algorithms for text categorization and also solved issue of high dimensional database that still persist in the table-based matching algorithm. In the suggested methodology, a document or documents are encrypted into table, as an alternative to the arithmetical vector. The dimensionality reduction is addressed by using the naturally inspired genetic algorithm after the learning. This algorithm minimized the size of the each document in the categorical profile. The implementation of the suggested methodology was authenticated in section 5 by means of the test beds such as 20NewsGroups. The experimental results of the proposed approaches showed that the performance evaluation is good compared to the other existing algorithms.

REFERENCES

[1] Taeho Jo, "Normalized table-matching algorithm as approach to text categorization", Soft Computing, Volume 19, Issue 4, pp 839-849, April 2015, Springer.

[2] Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., Wang, Z., 2007. A novel feature selection algorithm for text categorization. Expert Syst. Appl., 33(1):1-5.

[3] Zhou, B., Yao, Y.Y., Luo, J., 2010. A Three-Way Decision Approach to Email Spam Filtering. Proc. 23rd Canadian Conf. on Artificial Intelligence, p.28-39.

[4] Qi, X.G., Davison, B.D., 2009. Web page classification: features and algorithms. ACM Comput. Surv., 41(2):12-42.

[5] Coussement, K., van den Poel, D., 2008. Integrating the voice of customers through call center emails into a decision support system for churn prediction. Inf. Manag., 45(3):

[6] Wang, S.G., Li, D.Y., Song, X.L., Wei, Y.J., Li, H.X., 2011. A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification. Expert Syst. Appl., 38(7):8696-8702.

[7] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional Word Clusters vs. Words for Text Categorization," Journal of Machine Learning Research, vol. 3, pp. 1183-1208, 2003.

[8] Lee, L.H., Isa, D., Choo, W.O., Chue, W.Y., 2012. High relevance keyword extraction facility for Bayesian text classification on different domains of varying characteristic. Expert Syst. Appl., 39(1):1147-1155.

[9] de Souza, A.F., Pedroni, F., Oliveira, E., Ciarelli, P.M., Henrique, W.F., Veronese, L., Badue, C., 2009. Automated multi-label text categorization with VG-RAM weightless neural networks. Neuro computing, 72(10-12):2209-221

[10] Miao, Y.Q., Kamel, M., 2011. Pairwise optimized Rocchio algorithm for text categorization. Pattern Recogn. Lett., 32(2):375-382.

[11] Wang, B.K., Huang, Y.F., Yang, W.X., Li, X., 2012. Short text classification based on strong feature thesaurus. J. Zhejiang Univ.-Sci C (Comput. & Electron.), 13(9):649-659.

[12] Chen, E.H., Lin, Y.G., Xiong, H., Luo, Q.M., Ma, H.P., 2011. Exploiting probabilistic topic models to improve text categorization under class imbalance. *Inf. Process. Manag.*, 47(2):202-214.

[13] Wang Xiaoping, Cao Liming. Genetic algorithms-theory, application and implementation [M] . Xi'an Xi'an Jiaotong University Press. 2002.

[14] E. F. Combarro, E. Monta~n´es, IJ. Ranilla, and R. Mones, "Introducing a Family of Linear Measures for Feature Selection in Text Categorization," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 9, pp. 1223-1232, 2005.

[15] H. H. Hsu, C. W. Hsieh, "Feature Selection via Correlation Coefficient Clustering," Journal of Software, vol. 5, no. 12, pp. 1371-1377, 2010.

[16] J. Y. Jiang, R. J. Liou, S. J. Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 3, pp. 335-349, 2011.

[17] Y. X. Lin and B. C. Chien, "A Discriminant based Document Analysis for Text Classification," in proceedings of International Computer Symposium, Workshop of Artificial Intelligence, Knowledge Discovery, and Fuzzy Systems, Tainan, Taiwan, pp. 594-599, Dec. 16-18, 2010.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[18] Y. X. Lin and B. C. Chien, "Efficient Feature Reduction for High-Precision Text Classification," in Proceedings of the 2011 National Computer Symposium, Chia-Yi, Taiwan, pp. 36-45, 2011.

[19] J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi, and Z. Chen, "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 3, pp. 320-331, 2006.

[20] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C., 2002, Text Classification with String Kernels, Journal of Machine Learning Research, Vol 2, No 2, pp. 419-444.

[21] Banerjee, A., Dhillon, I., Ghosh, J., and Sra S., "Generative model-based clustering of directional data", The Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp19-28.

[22] Taeho Jo and Geun-Sik Jo, "List based Matching Algorithm for Classifying News Articles in NewsPage.com", IEEE International Conference on System of Systems Engineering, pp. 1-5, 2008, IEEE.

[23] Taeho Jo, "Using Semantic Similarity Matrix for Defining Operations involved in NTSO for Clustering 20NewsGroups", IEEE Congress on Evolutionary Computation (CEC), pp. 1-6, July 2010, IEEE.

[24] Taeho Jo, "Table Based Single Pass Algorithm for Text Clustering by Encoding Documents into Tables", IEEE International Workshop on Semantic Computing and Applications, 2008, IEEE.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

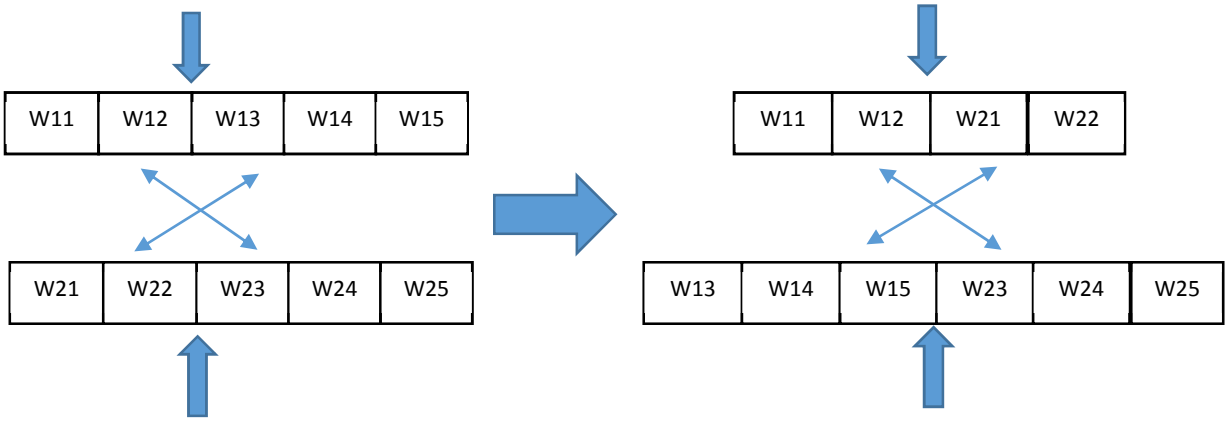

Fig 2: Crossover Operation For Text Categorization