Classifying Content Mode of Organizational Texts Using Simple Neural and Neuro-Fuzzy Approaches

Maryam Tayefeh Mahmoudi[1,2], Babak N. Araabi[2], Kambiz Badie[1], Nafiseh Forouzideh[3]
1: Knowledge Engineering & Intelligent Systems Group
IT Research Faculty, Iran Telecom Research Center
Tehran, Iran
Emails: {Mahmodi, k_badie}@itrc.ac.ir
2: Control and Intelligent Processing Centre of Excellence,
School of Electrical and Computer Eng., University of Tehran,
Tehran, Iran
Emails: {tayefeh, araabi }@ut.ac.ir
3: Kish Intl. Campus, University of Tehran,
Kish, Iran
Email: n.forouzideh@gmail.com

***Abstract*—In this paper, we present simple neural and neuro-fuzzy approaches to classify the mode of a text's content which is organized for helping users with their organizational tasks. In this regard, 7 major features were chosen as inputs for our suggested approaches. 3 nominal values L, M, and H were used as the possible values for each feature. Results of experimentation on a dataset including 540 data show the fact that the Takagi-Sugeno as a neuro-fuzzy approach using lolimot learning algorithm, performs better compared to multi-layer perceptron and radial basis function as simple neural approaches. Due to the high performance of this approach, it is expected to be successfully applicable to a wide range of content mode classification issues in decision support environment.**

Keywords- text classification; neural network; neuro-fuzzy approach; organizational task; content mode.

I. INTRODUCTION

In recent years, text mining has been widely used to extract the significant information from a text, among which extracting facts or regularities as well as the focal points are mentionable [1, 2, 3, 4]. An important point in this concern is the type(s) or class (es) to which a text or parts of a text may belong to. This has made classification one of prime issues in text mining [5, 6]. One major aspect in text classification is to identify the type of a text's content, e.g., its mode/style, its peculiarities/ characteristics, the category it belongs to, as well as the peculiarities of the environment within which it has been prepared. Pattern recognition techniques have a wide range of applications in this issue. Due to the distributed characteristics of a text, i.e., the fact that its mode/style may exhibit itself in an aggregation of a variety of considerations in its different parts/ components, non-symbolic classification methods equipped with logic of uncertainty handling like probabilistic and fuzzy logic are expected to be particularly workable in this regard.

Based on the above point, in this paper, we present an approach for classifying the mode of a text's content using neuro-fuzzy techniques [7, 8]. Due to the significance of comprehensive contents in making efficient decisions in organizations, the content mode considered in our approach is the type of an organizational task with regard to which texts have been organized.

The structure of the paper is as follows. Section II represents the existing approaches to text classification systems, while the emphasis of Section III is on the proposed approach including "the architecture of the proposed approach", "feature selection", and "experimental results" as well. Concluding remarks is also presented in Section IV.

II. EXISTING APPROACHES TO TEXT CLASSIFICATION SYSTEMS

Text classification can be defined as assigning texts to a predefined set of categories, which is used in situations where classes of texts/contents are labeled and include specific features. In this regard, from the viewpoints of similarity and regularity in features, the input content/text is supposed to be finally classifiable in terms of some pre-defined classes that can be significant in some sense. Within this context, classification can be performed based on the type of content, subject/issue, qualification level and style of content/text and even its authors' specifications. Text classification can also ease the organization of increasing textual information, in particular Web pages and other electronic form of documents [9]. It usually consists of two parts: feature selector and text classifier.

 ISBN: 978-1-61208-110-6

Feature selector selects the features which are essential to classifying the text's content; in terms of a feature vector. The classifier then assigns the feature vector to the appropriate class (es). Researches indicate that many techniques can be used in feature selection to improve accuracy as well as to reduce the dimensions of the feature vector and thus reduce the time for computation. Feature selection mostly adopts various assessment functions such as document frequency, information gain, mutual information, and statistics (CHI) to calculate the weights [10]. Many classifiers have been applied to classify texts, including Naïve Bayes [11], decision trees [12], k-nearest neighborhood [13], linear discriminate analysis (LDA) [14], logistic regression [15], neural networks [6], support vector machines [16], rule learning algorithms [17], relevance feedbacks [18], etc. Several kinds of competitive networks are used in text classification, including learning vector quantization (LVQ) and self-organizing maps (SOM) network. These two are both variants of the basic unsupervised competitive network. Besides, back propagation (BP) and radial basis function (RBF) networks are two successful examples for classification. There also exist some other statistical approaches for modeling a document for text classification like LSA, pLSA, and LDA [19].

Some special classification methods are also available for specific purposes, like Rocchio, which is for text classification in information retrieval [5] and independent component analysis (ICA), which was developed for the blind signal decomposition and recently used for selecting the mutually independent features of a document [20]. Text representation may also have a significant role in classifying texts with several features [21]. A series of experiments on text classification using multi-word features have also been done [22]. Meanwhile, web text classification has also been introduced as one of the major activities in this field [23].

III. THE PROPOSED APPROACH

Due to the distributed characteristics of a text and the fact that its mode/style may exhibit itself in an aggregation of several considerations in its different parts/components, non-symbolic classification methods equipped with logic of uncertainty handling are expected to function more efficiently.

Based on the above point, in this paper, we present an approach for classifying the mode of a text's content using neuro-fuzzy techniques. The mode of text's content considered in our approach is the type of an organizational task with regard to which the text has been organized using the dataset that has been prepared on the basis of the existing technical reports at a research institute. It is interesting to see that these tasks are equally being used by a wide range of knowledge workers (researchers, innovators, developers, planners, analyzers, etc.) in an organization to disseminate results of their works in terms of appropriate contents. Some of the major tasks important for an organization are: Planning/Scheduling, Research, Innovation, Development/ optimization/ Improvement, Education/ Promotion, Analysis/ Assessment/ Assurance, and Guidance, Justification.

In this paper, six of these tasks Research, Development/Planning, General Learning, Justification, Innovation and Analysis/ Assessment, etc are considered as the output classes.

A. The Architecture of the Proposed Approach

In this paper, the focus is on classification of a text's content using neuro-fuzzy approach. Neuro–fuzzy approaches in general and neuro-fuzzy networks in particular are fuzzy models that are not solely designed by expert knowledge but are at least partly learned from experiential data. If no a-priori knowledge is available, the application of a fuzzy model does not make any sense from the model accuracy point of view. However, if accuracy is not the only ultimate goal and instead an understanding of the functioning of the process is desired, then fuzzy models are the best choice [7]. In this respect, features of each functionality of a text's content can be identified and valued. These functionalities are considered to be the 6 classes of organizational tasks discussed above. In this regard, 27 features defined, out of which 7 major features have been chosen as inputs for our neural and neuro- fuzzy approaches. The important features are: "General Background", "Existing viewpoints", "Key issue", "Proposed approach realization/ implementation", "Validation/Verification", "Comparative analysis & capability interpretation", "Conclusion & prospect anticipation". The values of each of the features have been determined by experts. For instance, in a general learning content, for each feature of "General Background", "Existing viewpoints", "Key issue", the nominal values of "L" (Low), "H" (High), and "M" (Medium) have been determined. Detailed information about the features, their values and output classes are represented comprehensively in the next section. Figure 1 illustrates the overview of our proposed classification system.

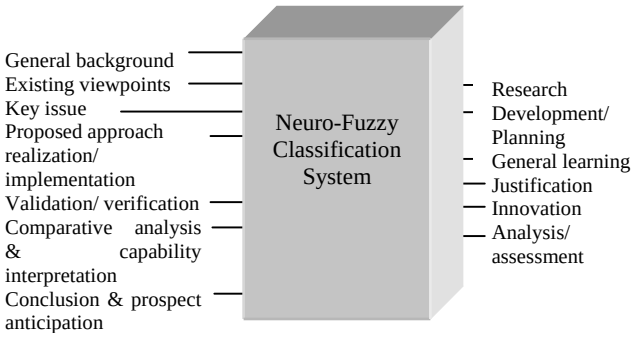

Figure 1. General view on the proposed classification system

In this paper, for classification purpose, both simple neural and neuro-fuzzy techniques have been considered. In this respect, multi-layer perceptron (MLP) and radial basis function (RBF) are implemented for simple neural and Takagi-Sugeno with Lolimot learning algorithm is implemented for neuro-fuzzy classification.

 ISBN: 978-1-61208-110-6

The experiments have been done on a dataset with 540 data, which have been prepared on the basis of the existing technical reports.

B. Feature Selection

With respect to mining issues, classifying the patterns existing in a database is of great importance, and due to this, selecting appropriate features for classification would also be significant. The high number of features in feature vector, makes in practice some difficulties when neural net is used as the classifier. In this respect, the major informative and uncorrelated features should be selected for classification [24].

In our approach, the appropriate features for classifying text's content are identified on the basis of the expert's idea and the existing approaches [25] as well. Within 27 previously identified features [25], 7 major features have been considered in this paper. Table 1 illustrates these features with their prospected values. It is obvious that, these features have been realized to be consistent for a wide range of contents which are to be created for helping users with their tasks in organizations, as discussed in the beginning of the section.

Obviously, based on the type of a task, a limited number of the labels and the corresponding sub-labels may be activated. Nominal values "L" (standing for Low), "M" (standing for Medium), and "H" (standing for High) associated with the labels of key segments indicate the extent according to which linguistically significant notions such as "What", "Who", "Whom", "Where", "Which", "When", "How", and "Why", can be addressed to create a petit content for each key segment in the content. This is done by the nominal values pre-agreed for each task, to show to what extent linguistically significant notions like "What", "Which", "Where", "When", "Whom", "Who", "Why", and "How" should be addressed [25].

Taking this point into account, the feature vector of input content is structured based on the afore-mentioned features and the nominal values (Table 1).

TABLE I. INPUT FEATURES AND OUTPUT CLASSES OF PROPOSED SYSTEM

Output Classes / Input Content Features	Research	Development / Planning	General Learning	Justification	Innovation	Analysis/ Assessment
General Background	H	M	L	L	M	L
Existing viewpoints	H	M	H	L	M	L
Key issue	H	M	M	M	M	M
Proposed approach realization/ implementation	H	M	L	M	L	M
Validation/ Verification	H	M	L	M	L	H
Comparative analysis & capability interpretation	H	M	L	L	L	L
Conclusion & prospect anticipation	H	H	L	L	L	L

Taking this point into account, the dataset used in this research would include the data from text/ content's labels that belong to 6 classes. It contains 540 samples with 7 attributes. After normalizing the input data and making the test and train data, classification would start.

C. Experimental Results

Simple neural approaches are used when no particular emphasis is made on the status of uncertainty in the related data, while neuro-fuzzy approaches are used to consider such a status of uncertainty. In the paper, we consider both of the approaches to show that uncertainty of the information in content is a matter which can not be disregarded [7, 18, 26].

1) Classification using MLP

In this respect, a feed forward MLP has been used with 1 hidden layer and a variation of hidden neurons. The optimal number of neurons in this respect was found to be 20. As we have 6 output classes, the binary forms of these classes would be as follows:

Output1/Class1 -> [0 0 1]
Output2/Class2 -> [0 1 0]
Output3/Class3 -> [0 1 1]
Output4/Class4 -> [1 0 0]
Output5/Class5 -> [1 0 1]
Output6/Class6 -> [1 1 0]

It is to be noted that if we divide the network into sub networks, the learning rate increases. In this regard, three networks of binary form of output classes are trained with normalized input data. The specifications of these three binary networks are as follows:

Number of neurons: 20; Train parameter epochs: 100; DivideParam.trainRatio = 0.7; DivideParam.testRatio = 0.15; DivideParam.valRatio = 0.15; Train Param. max_ fail = 30;

After training each network separately, the total network output is computed and is transformed from binary into decimal to have checked its status of belonging to the existing classes. Reconstructing test data for outputs and comparing the real classes with the network outputs yields realization of the whole classification process. The status of networks outputs are as follows:

1) First network: the best performance of validation, with least MSE is 0.029983 at epoch 6. The regression status shows 0.99, 0.94 and 0.89 learning respectively for the training, the test, and the validation data. Taking this point into account, 0.97 learning will be the result of the first experimentation.

2) Second network: the best performance of validation, with least MSE is 0.074266 at epoch 9. The regression status shows 0.94, 0.88 and 0.74 learning respectively for the training, the test, and the validation data. Taking this point into account, 0.90 learning will be the result of the second experimentation.

3) Third network: the best performance of validation, with least MSE is 0.001768 at epoch 99. The regression

 ISBN: 978-1-61208-110-6

status shows 0.999, 0.943 and 0.996 learning respectively for the training, the test, and the validation data. Taking this point into account, 0.9907 learning will be the result of the first experimentation.

As a result, it can be mentioned that the 3rd network learns totally better than the others networks with the rate of 99%, although it needs more epoch to reach the least MSE.

Results of the classification experiments on all the 540 data of the input dataset is illustrated in Figure 2.

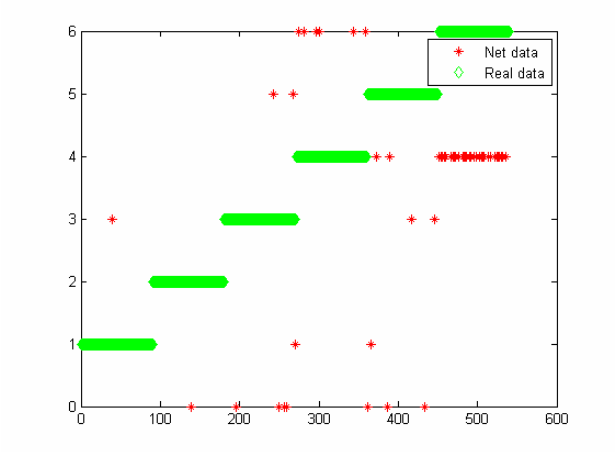

Figure 2. The classification result on all the input data

As it is seen, 57 inputs among 540 were false classified. The resultant total classification error is 10.55%, while the classification accuracy is 89.44%.

The experiments on the test data, reveals that 10 data were classified falsely. The classification error on the test data is 12.34% while the accuracy of the correct classification on test data is: 87.65%.

2) Classification using RBF

Another neural method for identification which has been used in this experiment is RBF whose basis function is Gaussian.

After normalizing all the inputs, the training and the test data are structured and all the three outputs are then computed. Based on the following conditions, the RBF is trained and tested on all data for the three networks:
goal = 0; spread = 1; MaxNeurons = 30; displayInterval = 2.
The status of outputs of the networks is as follows:

- First network: The best performance of NRBF is 0.0207203, considering Goal=0.
- Second network: The best performance of NRBF is 0.0585283 considering the Goal=0.
- Third network: The best performance of NRBF is 0.0136336, considering Goal=0.

As a result, it can be mentioned that the 3rd network totally learns better than the other networks with the rate of 85%.

Results of experiments on all the 540 data of input dataset are illustrated in figure 3.

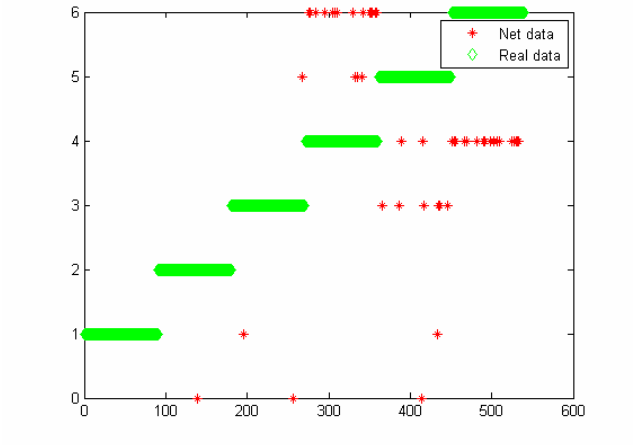

Figure 3. Classification result for all the data using NRBF

As it is seen, 48 inputs among 540 were false classified. The resultant total classification error is 8.88%, while the classification accuracy is 91.11%.

The experiments on the test data, reveals that 10 data were classified falsely. The classification error on the test data is 12.5% while the accuracy of the correct classification on test data is: 87.5%.

3) Classification Using Takagi-Sugeno with Lolimot

The neuro-fuzzy method applied for classification is Takagi-Sugeno with Lolimot learning algorithm [27]. As it is known, this method starts with an initial model, finds worst linear language model (LLM), checks all the divisions, finds best division and finally tests for convergence [7, 28].
Considerations for this experiment are as follows: smoothing factor (alpha) =1/3, mse_goal=1e-4 and reg_coef=0. Training the same three binary networks as previous parts, with max 30 neurons reveals that, the appropriate numbers of neurons for them respectively are: 6, 29, and 12.

As a classification result, it is to be noted that among 81 test data as input, 7 were classified falsely. Taking this point into account, the classification error was realized to be 8.64% and Lolimot was therefore able to distinguish 91.36% correct classes.

Figure 4 illustrates the classification status for both training and test data (540 input data). As it is seen, 23 data were classified falsely. In this regard, the total error of the network was realized to be 4, 26%.

 ISBN: 978-1-61208-110-6

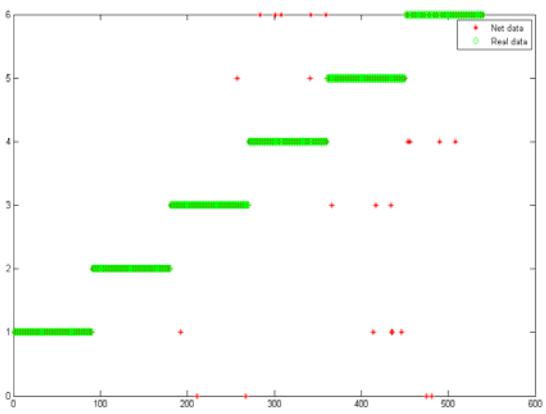

Figure 4. The classification of all the input data using TSK using lolimot

TABLE II. THE CLASSIFICATION RESULTS ON THE GIVEN DATASET FOR MLP, RBF AND TSK USING LOLIMOT

	MLP	RBF	Takagi-sugeno using lolimot
Appropriate No. of neurons for output1	20	30	6
Appropriate No. of neurons for output2	20	30	29
Appropriate No. of neurons for output3	20	30	12
Correct classification rate for Test data	% 87.65	% 87.5	% 91.36
False classification rate for Test data	% 12.34	% 12.5	%8.64
Correct classification rate for Whole data	% 89.44	% 91.11	%95.74
False classification rate for Whole data	% 10.55	% 8.88	%4.26
No. of corrected classified on Test data	71/81	71/81	74/81
No. of corrected classified on whole data	483/540	492/540	517/540

Figure 5 shows the comparison between the classification rates respectively belonging to MLP, RBF and Takagi-Sugeno using Lolimot. As it is seen from the experimental results, Takagi-Sugeno using Lolimot has classified better on test data compared to MLP and RBF.

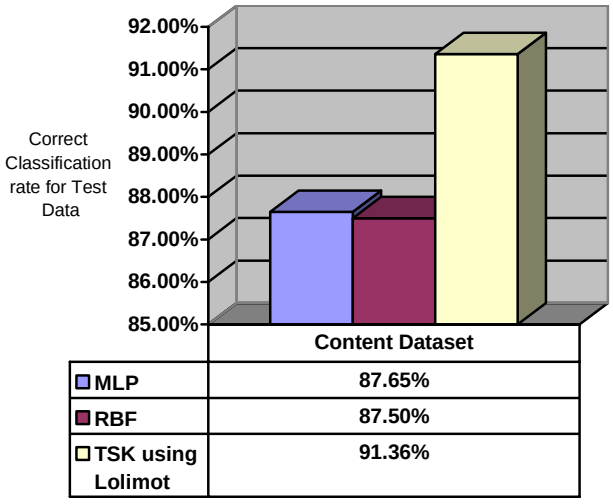

Figure 5. The comparison between Percent of corrected classification on Test data by MLP, RBF and TSK using Lolimot

The classification results on the whole data are illustrated in Figure 6. As it is seen, again TSK using Lolimot performs better than RBF, and RBF better than MLP.

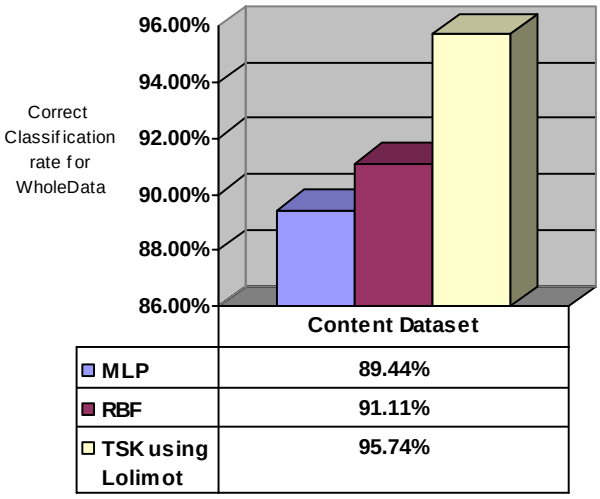

Figure 6. The comparison between Percent of corrected classification on whole data by MLP, RBF and TSK using Lolimot

As a conclusion, Takagi-Sugeno using lolimot learning algorithm, reveals better performance on the given dataset compared to the other mentioned algorithms. This is at first glance because a text's content has generally a multi-class or multi-modal nature, and thus due to its simultaneous affiliation to different classes (modes), classification approaches based on a sort of uncertainty handling logic can perform far better compared with those without such a basis. Moreover, the very peculiar ability of lolimot as a learning algorithm in speeding up the training procedure as well as incorporating with many kinds of prior knowledge (nominal

 ISBN: 978-1-61208-110-6

values for the input features in our case), and also its insensitivity toward curse of high dimensionality makes utilization of neuro-fuzzy approach more successful.

IV. CONCLUDING REMARKS

In this paper, the performance of multi-layer perceptron and radial basis function as simple neural approach, and Takagi-Sugeno with lolimot learning algorithm as neuro-fuzzy approach was evaluated for classifying the mode of a text's content, which is basically designed for helping users with their organizational tasks.

Experimental results on an initial dataset including the data belonging to 540 texts, demonstrate the fact that the Takagi-Sugeno with lolimot learning algorithm performs far better compared to simple neural approaches. This, as was discussed, is mainly due to ability of this approach in classifying the patterns of texts, which are somewhat multi-class or multi-modal in nature.

The approach presented in this paper can be particularly useful for organizing texts in decision support environments, where enriching the existing texts for supporting the human elements with their decisions (as the possible labels for content mode) is of particular significance.

REFERENCES

[1] D. S´anchez, M. J. Mart´ın-Bautista, I. Blanco, and C. Justicia de la Torre, "Text Knowledge Mining: An Alternative to Text Data Mining", 2008 IEEE Intl. Conf. on Data Mining , pp. 664-672.

[2] W. Wang, C. Wang, X. Cui, and A. Wang, "Fuzzy C-Means Text Clustering with Supervised Feature Selection," Fifth Intl. Conf. on Fuzzy Systems and Knowledge Discovery, 2008, Vol. 1, pp. 57-61.

[3] Y. Lu, S. Wang, S. Li, and C. Zhou, "Text Clustering via Particle Swarm Optimization", IEEE Conf. on Swarm Intelligence Symposium, (SIS '09), 2009, pp. 45-51.

[4] R. Li, J. Zheng and C. Pei, "Text Information Extraction Based on Genetic Algorithm and Hidden Markov Model" , First Intl. Workshop on Education Technology and Computer Science, Vol.1, 2009, pp.334-338.

[5] A. Danesh, B. Moshiri, and O. Fatemi, "Improve Text Classification Accuracy based on Classifier Fusion Methods", 10th Intl. Conf. on Information Fusion, 2007, pp. 1-6.

[6] Z. Wang, Y. He, and M. Jiang, "A Comparison among Three Neural Networks for Text Classification", IEEE Intl. Conf. on Signal Processing, 2006, pp. 1883-1886.

[7] O. Neles, "Nonlinear System Identification", Springer Pub., 2001.

[8] D. Kukolj, and E. mil Levi, "Identification of Complex Systems Based on Neural and Takagi–Sugeno Fuzzy Model", IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 34, No. 1, Feb. 2004.

[9] M. R. Islam and M. R. Islam, "An Effective Term Weighting Method Using Random Walk Model for Text Classification", 11th Intl. Conf. on Computer and Information Technology (ICCIT 2008), 2008, pp. 411 - 414.

[10] Z. T. Yu , L. Han, C.L. Mao , J. Y. Guo, X. Y. Meng , and Z. K. Zhang, "Study on the Construction of Domain Text Classification Model with the Help of Domain Knowledge", Seventh Intl. Conf. on Machine Learning and Cybernetics, 2008, pp. 2612 - 2617.

[11] P. Frasconi, G. Soda, and A. Vullo, " Text categorization for multi-page documents: a hybrid naive Bayes HMM approach". In Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (Roanoke, Virginia, United States). JCDL '01. ACM, New York, NY, 2001, pp. 11-20.

[12] R. E. Schapire and Y. Singer, "BoosTexter: a boosting-based system for text categorization", Machine Learning Journal, Vol. 39, No. 2/3, 2000, pp. 135–168.

[13] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Using *k*NN model for automatic text categorization" , Journal of Soft Computing - A Fusion of Foundations, Methodologies and Applications Vol. 10, No. 5, 2006, pp. 423-430.

[14] W. Li, L. Sun, Y. Feng and D. Zhang, "Information Retrieval Technology", Smoothing LDA Model for Text Categorization, Lecture Notes in Computer Science, Vol. 993, 2008, pp. 83-94.

[15] A. Genkin, D. Lewis, and D. Madigan, " Large-scale Bayesian logistic regression for text categorization", Technometrics Journal, Vol. 49, No. 3, 2007, pp. 291–304.

[16] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol. 34, No. 1, 2002, pp. 1–47.

[17] C. Apt´e, F.J. Damerau, and S.M. Weiss, „Automated learning of decision rules for text categorization", ACM Trans. on Inforation Systems, *Vol. 12*, No. 3, 1994, pp. 233–251.

[18] M. R. Azimi-Sadjadi, J. Salazar, S. Srinivasan, S. Sheedvash, "An Adaptable Connectionist Text Retrieval System With Relevance Feedback, IEEE Trans. on Neural Networks, Vol. 18, No. 6, Nov. 2007.

[19] J. Wang, X. Geng, K. Gao, and L. Li, "Study on Topic Evolution Based on Text Mining", Fifth Intl. Conf. on Fuzzy Systems and Knowledge Discovery ,Vol. 2, 2008, pp. 509-513.

[20] M. Hu, S. Wang, A. Wang, and L. Wang, "Feature Extraction Based on the Independent Component Analysis for Text Classification", Fifth Intl. Conf. on Fuzzy Systems and Knowledge Discovery, 2008, Vol. 2, pp. 296-300.

[21] J. Wang and Y. Zhou, "A Novel Text Representation Model for Text Classification", First Intl. Conf. on Intelligent Networks and Intelligent Systems, 2008, pp. 702-705.

[22] W. Zhang, T. Yoshida, and X. Tang, "Text classification using multi-word features", IEEE Intl.. Conf. on Systems, Man and Cybernetics, 2007, pp. 3519-3524.

[23] S. Yin, Y. Qiu, and J. Ge, "Research and Realization of Text Mining Algorithm on Web", Intl. Conf. on Computational Intelligence and Security Workshops, (CISW 2007), 2007, pp. 413-416.

[24] M. Rezaei, "Input Variable Selection in System Identification Application in Time Series Prediction", M.SC Thesis, University of Tehran, Feb 2008.

[25] K. Badie, M. Kharrat, M. T. Mahmoudi, M. S. Mirian, S. Babazadeh, and T. M. Ghazi, "Creating Contents based on Inter-play Between the Ontologies of Content's Key Segments and Problem Context", The First Intl. Conf. on Creative Content Technologies (CONTENT 2009), 2009, pp. 626-631.

[26] R. D. Goyal, "Knowledge Based Neural Network for Text Classification", 2007 IEEE Intl. Conf. on Granular Computing, 2007.

[27] L. Germán , O. Massa, L. Corbalán, L. Lanzarini, and A. De Giusti, "Evolving Fuzzy Systems A new strategy for rule semantics preservation", http://citeseerx.ist.psu.edu/viewdoc/ summary, doi=10.1.1.87.9948.

[28] J. Rezaie, B. Moshiri, and B. N. Araabi, "Distributed Estimation Fusion with Global Track Feedback Using a Modified LOLIMOT Algorithm", SICE Annual Conf. 2007, pp. 2966-2973.

 ISBN: 978-1-61208-110-6