Improvement in Automatic Classification of Persian Documents by Means of Support Vector Machine and Representative Vector

Jafari Ashkan[1], Izadi Hamed[2], Hossennejad Mihan[1]
[1]Computer Science Department
Islamic Azad University Bushehr Branch
Iran
[2]Computer Science Department
University of Applied Science and Technology Abadeh Branch
Iran
{Ajafari35, haamedizadi, mihan.hossennejad}@gmail.com

ABSTRACT: *Representative Vector is a kind of Vector which includes related words and the degree of their relationships. In this paper the effect of using this kind of Vector on automatic classification of Persian documents is examined. In this method, preprocessed documents, extra words as well as word stems are at first found. Next, through one of the known ways, some features are extracted for each category. Then, the Representative Vector, which is made based on the elicited features, leads to some more detailed words which are better Representatives for each category. Findings of the experiments show that Precision and Recall can be increased significantly by extra words omission and addition of few words in the Representative Vectors as well as the use of a famous classification model like Support Vector Machine (SVM).*

Keywords: Documents Classification, Representative Vector, Stemming, Support Vector Machine

Received: 19 March 2011, Revised 28 April 2011, Accepted 1 May 2011

1. Introduction

As information is producing increasingly, pressing need to classify it in order to optimize information retrieval is highlighted. Finding necessary information is only possible through searching keywords by search engines. Scientists usually find their required information easily through reading valid journals related to their scientific fields. This is because most of the times a person who is searching some information doesn't know a specific definition about what he needs and can not choose a certain keyword based on which he can search. Therefore, people can better find their necessary information through paging books. When information is classified topically, every specialist can get some necessary information easily by searching information related to their fields and will not waste their time searching a lot of unrelated information and retrieved documents [1]. Here, classification of digital resources seems vital. Unless digital resources classify, because they are absent physically it looks they have lost. The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. The resources of unstructured and semi structured information include the world wide web, Governmental electronic repositories, news articles, biological database, chat rooms, digital libraries, online forums, electronic mail and blog repositories. Therefore, proper classification and knowledge discovery form these resources are an important area for research.

Natural Language Processing (NLP), data mining, and machine learning techniques work together to automatically classify and

discover patterns from the electronic documents. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization. However how these documented can be properly annotated, presented and classified. So it consists of several challenges, like proper annotation to the documents, appropriate document representation, dimensionality reduction to handle algorithmic issues [2], and an appropriate classifier function to obtain good generalization and avoid over-fitting. Extraction, integration and classification of electronic documents from different sources and knowledge discovery from these documents are important for the research communities.

Today the web is the main source for the text documents, the amount of textual data available to us is consistently increasing, and approximately 80% of the information of an organization is stored in unstructured textual format [3], in the form of reports, email, views and news etc. the [4], shows that approximately 90% of the worlds data is held in unstructured formats, so information intensive business processes demand that we transcend from simple document retrieval to knowledge discovery. The need of automatically retrieval of useful knowledge from the huge amount of textual data in order to assist the human analysis is fully apparent [5]. This paper examines automatic classification in Persian texts by means of Representative Vectors and Support Vector Machine. Finally, findings of the research on some data texts are presented.

1.1 Related Works

Market trend based on the content of the online news articles, sentiments, and events is an emerging topic for research in data mining and text mining community [6]. For these purpose state-of-the-art approaches to text classifications are presented in [7], in which three problems were discussed: documents representation, classifier construction and classifier evaluation. So constructing a data structure that can represent the documents, and constructing a classifier that can be used to predicate the class label of a document with high accuracy, are the key points in text classification.

Text classification is an important component in many informational management tasks, however with the explosive growth of the web data, algorithms that can improve the classification efficiency while maintaining accuracy, are highly desired [8].

Based on ant colony optimization a new feature selection algorithm is presented in [9], to improve the text categorization. Also in [10] the authors introduced a new weighting method based on statistical estimation of the importance of a word categorization problem.

The authors in [11] focused on the document representation techniques and demonstrate that the choice of document representation has a profound impact on the quality of the classifier. They used the centroid-based text classifier, which is a simple and robust text classification scheme, and compare four different types of document representations: Ngrams, Single terms, phrases and RDR which is a logic-based documents representation. The N-gram is a string-based representation with no linguistic processing. The Single term approach is based on words with minimum linguistic processing. The phrase approach is based on linguistically formed phrases and single words. The RDR is based on linguistic processing and representing documents as a set of logical predicates. In [12] the authors present significantly more efficient indexing and classification of large document repositories, e.g. to support information retrieval over all enterprise file servers with frequent file updates.

Document classification has many uses such as, automatic question answering systems [13], information filtering, unimportant e-mail classification and other related areas [14]. In [15], a new technique based on ontology is offered for classification. The authors in [16] propose a Poisson Naïve Bayes text classification model with weight enhancing method, and shows that the new model assumes that a document is generated by a multivariate Poisson model. They suggest per-document term frequency normalization to estimate the Poisson parameter, while the traditional multinomial classifier estimates its parameters by considering all the training documents as a unique huge training document.

In [17], some results about automatic Persian text classification by indexing 4-gram and 3-gram measures are shown. Investigation of different approaches about automatic text classification in a new environment is dealt with in [17]. In [18], Persian text classification through KNN algorithm and its phase copy is presented. In [19], age ranges in speakers can be determined by examination of the related features in their vocal cords. In order to do that and to make optimal distinction among different categories including several age ranges, SVM is used.

1.3 Innovation in this Paper

Here, Representative Vector is used to improve text classification in texts to which learning collection is possible, for example, a

collection of news which is elicited from different resources automatically and is not patterned. Thus, in classification process, use of one special resource for learning step can not present all words of that category. Structure of the article is as follows:

In the next section, we will introduce the aforesaid issue and its related terms. In Section 3, we elaborate on the proposed solution. In Section 4, the relevant experiments and their results are shown. The last Section includes conclusions and further researches.

2. Statement of the Problem

The objective of document classification is to find the best category for each document. We have a good collection which is labeled by people as train set. Some words are usually selected from the train set. This process is called characteristic elicitation. Characteristic elicitation includes a selection of subordinate words which are available in the train set. This is done in such a way that only the same words are used for the classification. This has two reasons: first, the speed of training and classification is increased because of reduction in words numbers in the set. Second, noise word omission leads to Precision increase. A noise word is a kind of word that causes error increase in classification after learning process, we use the acquired knowledge in a new data collection called test set [20].Our purpose is to expand the characteristic elicitation in order to use it in classification of new documents. The instrument that we use here to find the related words with selected characteristic and to improve classification results is named Representative Vector. Representative Vector includes all related words and the degree of their relationships [21]. In the next section, we will be familiar with this concept, its use and some ways of making it.

2.1 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are one of the discriminative classification methods which are commonly recognized to be more accurate. The SVM classification method is based on the Structural Risk Minimization principle from computational learning theory [22]. The idea of this principle is to find a hypothesis to guarantee the lowest true error. Besides, the SVM are well-founded that very open to theoretical understanding and analysis [23]. The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n-dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. The performance of the SVM classification remains unchanged if documents that do not belong to the Support Vectors are removed from the set of training data [24].

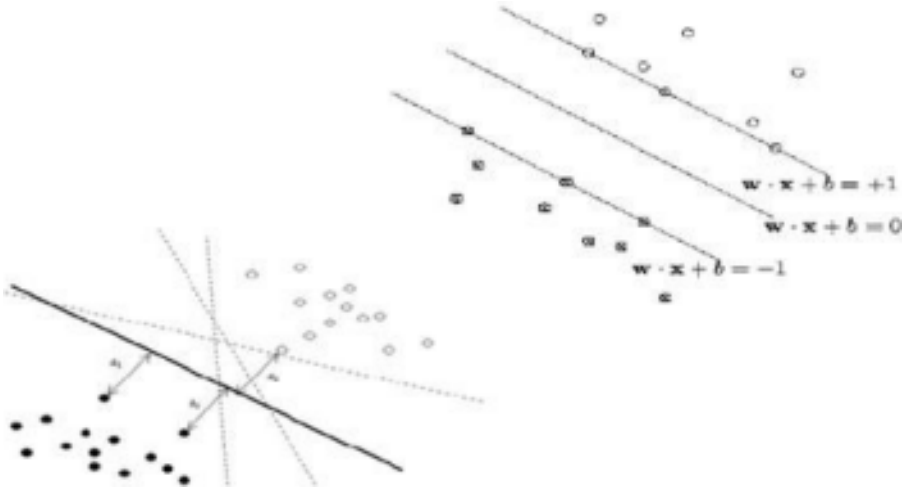

Figure 1. Illustration of optimal separating hyperplane, hyperplanes and Support Vectors[24]

The SVM classification method is outstanding from the others with its outstanding classification effectiveness [24] [25] [26] [27] [28] [29]. Furthermore, it can handle documents with high dimensional input space, and culls out most of the irrelevant features. However, the major drawback of the SVM is their relatively complex training and categorizing algorithms and also the high time and memory consumptions during training stage and classifying stage. Besides, confusions occur during the classification

tasks due to the documents could be a notated to several categories because of the similarity is typically calculated individually for each category [24]. So SVM is supervised learning method for classification to find out the linear separating hyperplane which maximize the margin, i.e., the optimal separating hyperplane (OSH) and maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which is "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier. The SVM is a best technique for the documents classification [30]. The authors in [29] implemented and measured the performance of the leading supervised and unsupervised approaches for multilingual text categorization; they selected support vector machines (SVM) as representative of supervised techniques as well as latent semantic indexing (LSI) and self-organizing maps (SOM) techniques for unsupervised methods for system implementation.

3. Solution Steps

General steps in the figure 3 are shown. At first, some features are elicited for each category. The corpus includes HAMSHAHRI news in which categories have already specified. In the next step, Representative Vectors are made for the elicited features. Representative Vectors present some words for features that have semantic relationship with that feature. We use these words to improve the collection of features for each category. In next sections we will explain these steps and demonstrate the results.

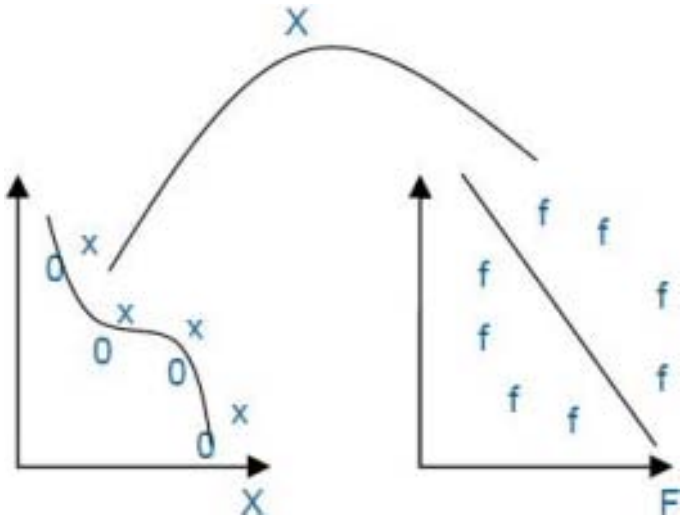

Figure 2. Mapping non linear input space onto high dimensional space[24]

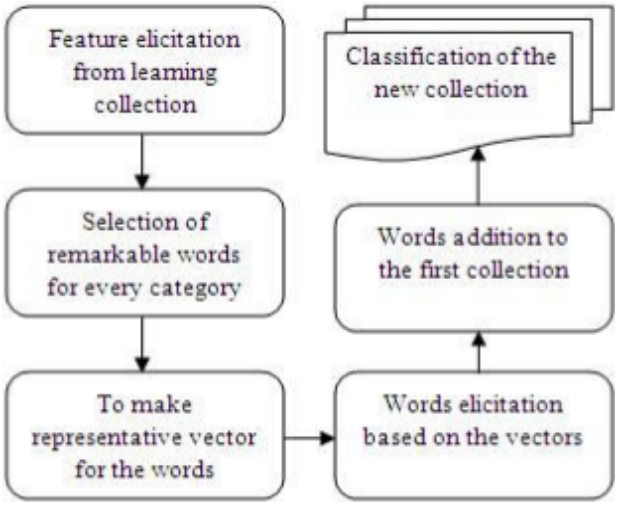

Figure 3. General steps in the offered algorithm

3.1 Feature Elicitation

For feature elicitation in every category, we have used the method MI. By means of MI we can realize to what extent presence or absence of a word in a document may inform us about a category [20]. Based on formula 1, every word and category is given a score. Then all words in each category should be ordered based on the weight that they gain in that category and the top ones should be chosen.

$$I(U,C) = \sum_{e_t \text{ in}\{1,0\}} \sum_{e_c \text{ in}\{1,0\}} (P(U = e_t, C = e_c) \times \log \frac{(P(U = e_t, C = e_c)}{(P(U = e_t) \times (C = e_c)'} \quad (1)$$

3.2 How to Make a Representative Vector

At first, we explain how to make Representative Vector. A series of documents used to make a Representative Vector are labeled C. the process starts with a word for which we want to make a Representative Vector. At first we assume that the given word is a kind of search. Then C is organized based on the search (according to one of retrieval methods like OKAPI). We select ten first documents and calculate a weight for each of the present words in these ten documents formula 2.

$$W_{d,t_i} = \frac{tf(t_{i,d}) \times (C_{doc} - df(t_i))}{\sum_i tf(t_{i,d}) \times C_{doc}} \quad (2)$$

W_{d,t_i} Is the weight of t_i in document *D*. Next, the final weight of every word in the collection of selected words for each category is calculated by formula 3.

$$W_{t_i} = \frac{\sum_{j=1}^{NoDocs} W'_{d_j,t_i}}{NoDocs} \quad (3)$$

W'_{d_j,t_i} Is the weight of t_i in document d_j. NoDocs is equal to the number of retrieved documents (i.e. 10).

3.2.1 Optimization Step

We have selected ten first words from among the best documents related to a word X so far. We have also specified the most important words from among all available words in these ten documents, which are labeled t_i. The acquired words by this way are those which have got high TF and IDF scores among documents related to word X. However, this high score may be because of factors other than semantic relationship between X and t_i. Thus, we add an extra phase to improve the relationship. All the steps are repeated for each t_i in this phase. If X is also present in the collection of words related to t_i, it is highly probable that there is a close relationship between X and t_i.

3.3 Use of Representative Vectors for Classification

Up to now, we elicited ten words for each category and made a Representative Vector for each of them. Now, we have ten Representative Vectors that each includes some related words with a concept. For example, in Table 1, ten elicited words form category اقتصاد ("economy") are shown in the left table by MI method. The right table demonstrates the Representative Vectors for the word بازار ("market") which are ordered based on their weights calculated in formula 3.

4. Experiments and Results

In this project, we have used the HAMSHAHRI corpus [31], as the train and test set. This corpus contains 160000 news between 1997 and 2002 years. We have only four categories including economy, politics, science and sport for the experiments. To make Representative Vectors ISNA corpus [21], which has more than 500 mega bytes of data, is used. Our experiment is comprised of two main phases. The steps done in the first phase are as follows:

The first step: extra words, prepositions and numbers are omitted for every category that is in the train set. Then by a simple evolution algorithm stems of the words are offered.

The second step: preprocessed documents are indexed and some features are elicited by MI. In the third step according to the

elicited features and the documents to the category, word-document matrix is made. In this matrix, columns indicate features and rows demonstrate the documents including the features. Each cell in this matrix indicate the weight TF-IDF of the feature in the mentioned document. The last column shows the category of documents. In the last step, we use Support Vector Machine and the train set is made. The test set is also made in this way. For the precise evaluation, first we use the test set while the documents are not preprocessed. Table 2, shows the details of the experiment.

دسته اقتصاد ("economy category")		بازار ("Market")		وزن ("weight")
	English Translation		English Translation	
سال	year	نفت	petrol	0.87320001
گزارش	report	دلار	dollar	0.83331113
افزایش	increase	جهانی	universal	0.70001112
ایران	Iran	قیمت	cost	0.68998710
درصد	percentage	طلا	gold	0.66663331
تولید	production	بانک	bank	0.57783333
اقتصادی	economical	فروش	sale	0.49999877
توسعه	extension	تورم	inflation	0.45321111
برنامه	program	سکه	coin	0.44398900
بازار	market	بشکه	barrel	0.41116111

Table 1. The left table: elicited words from economy category, the right table Representative Vector of the word بازار ("market")

In the next step, we preprocess the test set and omit extra words, prepositions and numbers and find the stems of the words. Table 3, shows the details of the experiment.

As we expect, preprocessing ways improve Precision and Recall in SVM. In the second phase of the experiments, all the steps done in the previous phase are repeated. Only with the difference that this time, we have an additional step called "how to make Representative Vector". For each elicited feature during this step, we make a Representative Vector MI, then, we add a few words to the collection of available features which are better Representatives for that feature. Table 4, demonstrates the effect of Representative Vector on Precision and Recall of classification by Support Vector Machine.

As we expect, preprocessing ways improve Precision and Recall in SVM. In the second phase of the experiments, all the steps done in the previous phase are repeated. Only with the difference that this time, we have an additional step called "how to make Representative Vector". For each elicited feature during this step, we make a Representative Vector MI, then, we add a few words to the collection of available features which are better Representatives for that feature. Table 4, demonstrates the effect of Representative Vector on Precision and Recall of classification by Support Vector Machine.

Category	Recall	Precision
Economy	0.62	0.939
Politics	0.49	0.7
Science	0.55	0.797
Sport	0.92	0.472
Average	64.5%	72.7%

Table 2. Precision and Recall of SVM Classifier While the Documents are not Preprocessed

Category	Recall	Precision
Economy	0.75	0.852
Politics	0.61	0.685
Science	0.84	0.808
Sport	0.82	0.689
Average	75.5%	76%

Table 3. Precision and Recall of SVM Classifier after Documents Preprocessing

Category	Recall	Precision
Economy	0.72	0.947
Politics	0.7	0.753
Science	0.82	0.854
Sport	0.89	0.659
Average	78.3%	80.3%

Table 4. The Impact of Representative Vector on Precision and Recall of SVM Classifier

5. Conclusion and Further Research

The main purpose of the paper, was to make a basis for the efficiency evaluation in Support Vector Machine. Document preprocessing plays an important role in the improvement of Precision and Recall of the model. After preprocessing these two measures increase to 3%, 11% Respectively. More over, the significant effect of the added words is that they improve the two

mentioned measures remarkably as 4%, 3% Respectively. In economy and science categories, after using the representative vector recall measures have decreased. Probably this is because some words like گزارش ("report") are so common in these two categories, that their presence can not help the improvement of the measures.

Now, we conclude that words increase in the train set as well as preprocessing improve the efficiency and precision of the classification. In further research we aim at examining the efficiency and Precision of other classifiers by this way. Furthermore, we offer an eclectic method for feature selection in order to improve efficiency and Precision of Support Vector Machine and other classifiers.

References

[1] Bina, B., Rahgozar, M., Dahmouyad, A. (2007). Automatic classification of Persian texts. 13th national conference computer forums Kish, IRAN.

[2] Dasgupta, A. (2007). Feature selection methods for text classification. *In:* Proc. of the 13th international conference on Knowledge discovery and data mining, p. 230 -239.

[3] Raghavan, P., Amer-Yahia, S., Gravano eds, L.(2004). Structure in Text Extraction and Exploitation. *In:* Proc. of the 7th international Workshop on the Web and Databases (WebDB), ACM Press, 67 (4) 240-255.

[4] Oracle corporation. (2008). URL: http:// WWW.oracle.com.

[5] Lynch, M. (2000). e-Business Analytics. Depth Report.

[6] Falinouss, P. (2007). Stock Trend Prediction using News Article's: a text mining approach. Master thesis.

[7] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), p. 1 – 47.

[8] Liu, H., Motoda, M. (2006). Feature Extraction, construction and selection: A Data Mining Perpective. Boston, Massachusetts(MA), Kluwer Academic Publishers.

[9] Sccuy, P., Mineanu, G. W. (2007). Beyoned TF-IDF weighting for text Categorization in the Vector Space Model. *In:* Proc. of 2nd international conference on recent advances in natural language processing, p. 241-248.

[10] Forman, G., Kirshenbaum, E. (2008). Extremely Fast Text Feature Extraction for Classification and Indexing. *In:* Proc. of 3nd international conference on recent machine learning, p. 341-348, Napa Valley California, USA.

[11] Keikha, M., Khonsari, A., Oroumchian, F. (2009). Rich document representation and classification: An analysis. *In:* Proc. of 2nd international conference on Knowledge-Based Systems, p. 67–71.

[12]Tam, V., Santoso, A., Setiono, R. (2002). A comparative study of centroid-based neighborhood-based and statistical approaches for effective document categorization. *In:* Proc. of the 16th International Conference on Pattern Recognition, p. 235–238.

[13] Moschitti, A. (2007). Answer filtering via text categorization in question answering systems. *In:* Proc. of 2nd international conference on recent advances in natural language processing, p. 241-248.

[14] Huang, Y. (2006). Support vector machines for text categorization based on latent semantic indexing. technical report, electrical and computer engineering department, Johns Hopkins university.

[15] Shang, W., Huang, H., Zhu, H. (2006). A Noval Feature Selection Algorithm for text catogorization. *Expert System with application*, 1, 1-5.

[16] Domingos, P., Pazzani, M. J. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, *Journal of Machine Learning*, 29 (2-3) 103-130.

[17] Esmail pour, M. (2006). Approaches and challenges category automatic classification of information resources in new environment, *Library and Information Update*, 10 (2).

[18] Basiri, M., Neimati, S., Ghasem aghayi, N. (2007). Compare Persian texts classified using KNN algorithm and the F-KNN and select features based on information gain and document frequency. *In:* Proc. Of 13th national conference computer forums Kish, IRAN.

[19] Homaun puor, M. M., Khosravi, M. H. (2007). To help determine the age range speaker sound using support vector machines, *In*: Proc. Of 13th national conference computer forums Kish, IRAN.

[20] Christopher, D., Manning, P. (2008). Introducion to Information Retrieval: Cambridge University Press.

[21] Amiri, H., AleAhmad, A. (2008). Keyword Suggestion Using Concept Graph Construction from Wikipedia Rich Documents. ECIR'08 Workshop on Exploiting Semantic Annotations for Information Retrieval, Glasgow.

[22] Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *In:* Proc. of 10th European Conference on Machine Learning, p. 137-142.

[23] Sahay, S. Support Vector Machines and Document Classification. URL: http://www-static.cc.gatech.edu/~ssahay/sauravsahay7001-2.pdf.

[24] McCallum, A., Nigam, K. (2003). A Comparison of Event Models for Naïve Bayes Text Classification. *Journal of Machine Learning Research*, 2 (3) 1265-1287.

[25] Chakrabarti, S., Roy, S., Soundalgekar, V. (2003). Fast and Accurate Text Classification via Multiple Linear Discriminant Projection. *International Journal on Very Large Data Bases* (VLDB), 1 (2) 170-185.

[26] Lin, Y. (1999). Support Vector Machines and the Bayes Rule in Classification. Technical Report, No.1014, Department of Statistics, University of Wiscousin, Madison.

[27] Sahay, S. Support Vector Machines and Document Classification. URL: http://www-static.cc.gatech.edu/~ssahay/sauravsahay7001-2.pdf.

[28] Yang, Y., Liu, X. (1999). A Re-examination of Text Categorization Methods. School of Computer Science, Carnegie Mellon University.

[29] Lee, C. H., Yang, H. C. (2009). Construction of supervised and unsupervised learning systems for multilingual text categorization. *In:* proc. of Expert Systems with Applications, p. 2400–2410.

[30] Isa, D., Kallimani, V. P. (2008). Using Self Organizing Map for Clustering of Text Documents. *In:* Proc. of Expert System with Applications, p. 44-50.

[31] Darrudi, E., Hejazi, M. R., Oroumchian, F. (2004). Assessment of a modern farsi corpus. The 2nd international workshop on information technology and its disciplines, Island, Iran, Feb.

Author biographies

Jafari Ashkan Obtained his B.SC. degree in computer software engineering in 2007 from Islamic Azad University of Meybod Branch and her M.SC. degree in computer software engineering in 2010 from Islamic Azad University of Zanjan Branch. His research interests include data mining, text mining, document classification, document clustering.

Hosseinnejad Mihan Obtained her B.SC. degree in computer software engineering in 2007 from Islamic Azad University of Khoei and her M.SC. degree in computer software engineering in 2010 from Islamic Azad University of Zanjan. Her research interests include information retrieval and text mining.

Izadi Hamed Obtained his B.SC. degree in computer software engineering in 2007 from Sepahan institute of higher education and her M.SC. degree in computer software engineering in 2010 from Islamic Azad University of Tehran North branch. His research interests include data mining and computer networks.