International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 1, No. 2, 2014, pp. 152-160.
ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

A Review of Researches on Persian Text Mining

Monireh Azimi Hemat[a], Fatemeh Shamsezzat[b]

[a] *Department of computer , Faculty of engineering ,Payame Noor University, P.O. Box 19395-3697, Tehran, IRAN*

[b] *Department of Computer Science, Faculty of Mathematics and Computer, Fasa University, Fasa, P.O. Box 74617-81189 , IRAN,*

Abstract

Due to the rapid growth of volume of text information, the existence of tools to organize, retrieve and extract useful knowledge from this data is essential. In this regard, no extensive research has been done on text mining methods so far. By using identification and detection of patterns, the text mining techniques extract useful information from unstructured text data. Text mining is known as text data mining, knowledge discovery in text and intelligent text analysis. This research has been done on Persian texts less than other language ones due to the complexity of Persian structure. In this paper, an overview of the main methods of text mining and their application in processing Persian language texts is provided. Although it is not possible to express all the methods and applications, this article can give general information to the reader and is a prelude to future work.

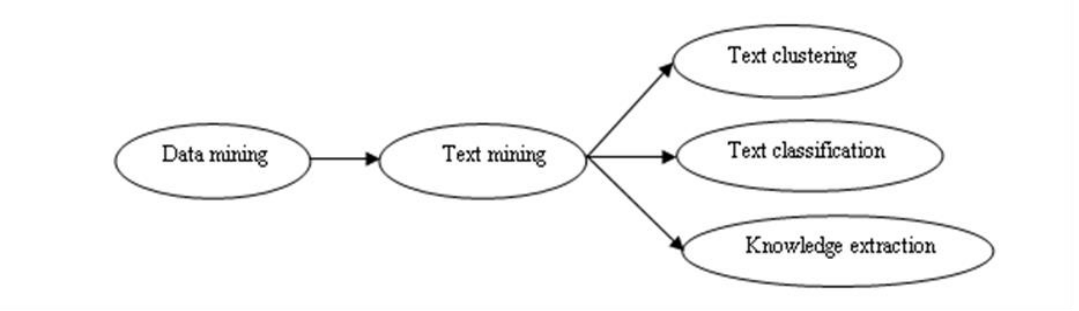
Keywords: Persian text mining, Persian text classification, Persian text clustering.

Introduction:

In today's world, information plays an important role. On the other hand, a significant portion of the information is stored in text databases. This huge amount of data needs to be processed so that useful knowledge can be extracted from them. Text mining techniques are used for comparing different texts, summarizing, ranking most important and most relevant texts, as well as finding patterns and behaviors among different texts. Text mining was proposed by Feldman and his colleague for the first time (Feldman and Dagan, 1995). Due to complex structural problems in Farsi, little research has been done in the field of text mining of Persian texts. In this article, after explaining the process of text mining, two general text mining methods are presented. Then, the problems of the text mining of Persian texts are explained and the application of text mining methods in Persian text processing is examined.

Text mining process:

Since a text a type of data, text mining is considered a subcategory of data mining. For this reason, data mining methods and algorithms can widely be used for text mining, with this difference that natural language processing methods should be applied. Therefore, generally text pre-processing tasks are done by using natural language processing methods and text processing by data mining methods. Figure 1 shows the relationship between text mining and data mining.



Shape 1- the relationship between text mining and data mining

aqakardan and Keyhan nejad (1391) defined the following steps for text-mining process:

- Document collection: to begin the process of text mining, the collection of documents takes place. The original documents that are raw materials are examined for extraction of knowledge.

- Pre-processing: this stage is performed in two phases of text and literary pre-processing. In the text pre-processing, tokenization, Stop-Word removal and stemming are done on primary documents (Hammouda and Kame, 2007). In the literary pre-processing, to increase the phrase information, Part-of-Speech tagging, Text Chunking, Word Sense Disambiguation (WSD) and Parsing are repeatedly applied (Hotho et al, 2005).

- Model Construction: In this stage, a document term- graph-based data model is built and their features are extracted.

- Pattern analysis: this stage is performed so that the similarities among the documents on the basis of important terms can be discovered. In this part, clustering is used to form document groups and extract key terms.

- High level knowledge: the process that is resulted by extracting key terms of clusters and determining subject identification of each clustering (Hotho et al, 2005).

Karami (1386) defined text mining as algorithm application and statistical methods to find new patterns and the following stages are presented for this process.

- Data retrieval: a process in which only documents that are detected relevant to the subject are set.

To find knowledge of the documentation, the statistical methods are used for automatic processing of textual data.

- Natural language processing: provides Possibility of understanding of written content via natural language (such as texts in English and Persian) by machine. These techniques have many applications in text mining, summarization and Farsi queries (Zhou et al, 2007).

- Data extraction: its aim is to get specific information from texts. This information results from mapping of natural language texts to a built and pre-defined display and are stored in the patterns similar to the databases for future use.

Persian Text mining

Most text mining systems are designed for English texts and are not utilizable for Persian texts. Due to the nature of the Persian language and the unavailability of a set including word root and common words, development of these systems to be applied for Persian texts is relatively difficult.

Morteza'i (1380) noted that some Persian language problems include:

- Diversity in scientific equations

- Recording Names

- Determining the word boundaries

- Joint writing

- Separate writing

- Plural types

- Different forms of writing

However, many articles are written in searching, summarizing, translating and spelling Persian. In the following section, a review is done on the methods of text mining as well as the work done in this area on Persian texts.

Text mining methods

In this section, two methods of classification and clustering and their applications in Persian text mining are described.

- Classification method

Text categorization assigns natural language texts to one or more predefined categories based on their content (Feldman and Sanger, 2007) (Chiang et al, 2008). There are many methods that are proposed to classify Persian texts.

(Bina et al., 1386) proposed a method for automatic classification of Persian texts. In this study, they stated that the text automatic classification includes Feature extraction phase in text and Learning phase (Cavnar and Trenkle, 1994). Also, the extraction of characteristics from Persian texts includes the phases of text preparation and indexing the documentation of index Weighting.

During the text preparation phase, the Persian text that includes the characters one after another is converted into display that is appropriate for the learning and classification algorithms. This process usually includes the following (Khreisat, 2004):

- To remove tags of html or xml

- To encode texts in utf-8

- To remove stop words and punctuation

- To find word stemming and remove prefixes and suffixes

In the phase of document indexing, the importance of each expression in all documents is determined (Aas And Eikvil, 1999). In this survey, document indexing method in the form of simple words and N-gram method are used (Cavnar and Trenkle, 1994).

The final phase is to weight extracting features and the following method is used for index weighting (Liao et al, 1997):

w= (log (freq))/((1+log(avg TF)))×log(n/filecounte+1)

Freq: The number of occurrence in the text

avg TV: average of phrase occurrence

n: number of texts

filecounte: the number of phrase in texts

The second step in the automatic classification of texts is learning that includes the definition of similarity measure, learning algorithm and assessment of efficiency.

Finally, this method of automatic classification of texts was tested on a set of 400 articles of Citizen Newspaper in seven classes. The best indexing method was 4-gram and the best value of n is obtained 20 at knn algorithm.

Sheikhi et al., (1391) presented and compared several methods to classify Persian texts. In their research, the stages for Persian literature classification have defined.

The first stage is text pre-processing that prepares the document for classification process. At this step, similar to many natural languages processing, tags and irrelevant words are removed. Indexing Stage uses a scheme to weight the phrases in the text.

The next step in the classification is choosing the appropriate feature in the document and the selection keys should be accurate and their number should not increase the computational complexity. After extracting keywords, different classification algorithms are applicable (Mohaqeqi & Soltanpoor, 2009). As mentioned before, feature selection plays a key role in the accuracy and efficiency of the system. In this research, the methods such as document frequency (DF), information gain (IG) and mutual information (MI) are presented (Yang & Pedersen, 1997).

To classify texts, these methods which are Naïve Bayes, KNN, decision tree and support vector machine are used (Mohaqeqi & Soltanpoor, 2009). In the end, after applying feature extraction methods and different classification of standard citizen data collection, the results show that support vector machine classifier has better performance.

sarai and Shahgholian (1389) proposed a method based on machine learning for classifying Persian texts. In machine learning method, there are training and testing phases.

To design machine learning algorithms, the type of training data, goal function, its display and an algorithm to learn it should be chosen carefully from training samples (Mitchell, 1997).

In this study, five phases of machine learning to classify text are as follows: pre-processing, feature production, feature selection, application of text mining patterns and interpretation and evaluation. These steps are explained briefly.

The purpose of the pre-processing phase is to find useful words and ignore useless words. After this phase, the texts are converted from the string of words to the vectors of words. There are no useless words in these vectors.

To obtain a proper input for text classification learning algorithms, these vectors and the feature production method are used. Respectively, the features of each word in the texts are found.

Given that the number of these features is still very high, some effective features should be chosen among them to classify the texts. To choose features in Persian text classification system, an adjustable threshold is used. In this article the threshold between 0 and 30 is selected.

KNN is one of simple and effective algorithms in text classification. The Basis of this algorithm is to compare test text to training texts and obtain the similarity between them (Tan, 2006).

In the last phase, to evaluate KNN algorithm, test texts that have been collected from various internet resources are used. The texts are located in seven categories and after the experiment, it is observed that which one of feature production formulas had better results for each category.

-Clustering method

One of the techniques in the field of text mining is clustering that investigates the similarity between data. This method is based on finding and grouping similar documents (Elahi, 1389).

In general two methods of clustering are hierarchical and partitioning. In the hierarchical method, after each document is considered as a cluster, by calculating the distance between pairs of clusters, each pair is merged with minimum distance. In partitioning methods, the data is divided into k partitions. Each of these partitions is considered as a cluster (Elavarasi, 2011).

One of the most famous partitioning methods is K-means algorithm which aims to minimize the mean square of the distances of the objects to their cluster centers. The following section is about some clustering applications in text mining of Persian texts.

Bazqandi et al., (1391) clustered the Persian sentences by using collective intelligence algorithms. Particle swarm algorithm (PSO) is an optimization algorithm that mimics the behavior of animal societies in processing society knowledge (Eberhart And Kennedy, 1995).

In this article, a novel method based on PSO is introduced for clustering sentences of a text. In this method, instead of using Euclidean and cosine distance, a new standard has been applied in calculating the sentences distance.

Bazqandi et al., specified the semantic relationship between words by using context-vector before they applied the PSO algorithm. The goal is to determine the relationship and the similarity of candidate words that are extracted from the text. Then, by defining criteria and functions that satisfy clustering features, convenient functions are extracted to optimize PSO algorithms. At the end, after algorithm evaluation, this method is introduced as one of the most appropriate methods in terms of efficiency.

Beheshtipour et al., (1392) used a statistical method of feature selection to remove superfluous and disordered words from data and improve clustering algorithm performance. They applied feature selection method that is a supervised algorithm in clustering algorithm of documents that is unsupervised.

In this paper, pre-processing including the removal of stop and unimportant words has taken on documents. Also, after examining and comparing several methods of feature selection, one that had the best results is selected:

1. After removing the stop words, rχ^2 (w) is calculated for each word in the document.

2. The words on the basis of the corresponding values rχ^2 (w) is sorted.

3. Some of the words that have the most rχ^2 (w) are selected.

Finally, K-means algorithm is improved by utilizing this feature selection method.

Discussion and conclusion

Text mining is considered as a growing and common area to discover knowledge from unstructured texts. Due to structural problems in Farsi, less research has been done in this area. Given the importance of

processing Persian text data in our country, this article reviews the work done in the field of text mining of Persian texts. Although the expression of all research conducted in this field is impossible, by giving a general view to the readers, this article can be a prelude for future research.

References:

Aqakardan, Ahmad, & Keyhan nejad, Mina., (2012) , A model to extract information from text documents based on text mining in the area of e-learning ،Journal of Information and Communication Technology,11, 47-54. (In Persian)

Karami, mahtab,(2007), The application of data mining and text mining analyzer tools in agility on healthcare organizations, Journal of Health Administration, 10(30), 15-20. (In Persian)

Mortezaei, Leila, (2001), Persian script and language issues in storage and information search, information Journal, 17. (In Persian)

Bina, bahare, & rahgozar, masood, & deh_mobed, azin, (2007), Automatic classification Persian texts, the Thirteenth National Conference of Computer Society of Iran, Kish Island, Iran. (In Persian)

Sheikh, Maryam, & AkbarPoor, shahin, & Farzan, Ali, (2012), Text mining to classify Persian literature, Iranian Conference on Electrical and Electronic Engineering, University of Gonabad. (In Persian)

Saraei, Mohamad hosein, & shahgholeian, Azar, (2010), Explore Persian texts based on the classification method, Science and Computer Engineering Journal, 1, 8-13. (In Persian)

Elahi, Abdalkarim, (2010), Improved clustering of documents based on a new similarity measure, Iran Data Mining Conference, Tehran. (In Persian)

Bazqandi, Mahdi, & Tadaion Tabrizi, Qamarnaz, & vafaei jahan, Majid,(2012), the first international conference on natural language and line processing, Semnan University. (In Persian)

Behishtipur, MR, & jafarei, Ali, & Javanbakht, Morteza, (2013), clustering algorithm Persian documents based on improved K-Means algorithm, the Seventh Scientific Conference of Command and Control, University of Imam Hossein. (In Persian)

Feldman, Ronen. & Dagan, Ido. (1995). KDT-Knowledge Discovery in Texts. Proc. of the First Int. Conf. on Knowledge Discovery KDD. 112-117.

Hammouda1, Khaled. & Kamel, Mohamed. (2007). Data Mining in e-Learning. E-Learning Networked Environments and Architectures Advanced Information and Knowledge Processing. Springer. Part IV. 374-404.

Hotho, Andreas. & N¨urnberger, Andreas. & Paaß, Gerhard.(2005). A Brief Survey of Text Mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology 20(1):19-62.

Zhou, Xuezhong.& Liu, Baoyan. & Wu, Zhaohui. & Feng, Yi. (2007). Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks. Arificial Intelligence in Medicine; 41(2): 87-104.

Cavnar , William B. & Trenkle, John M.(1994). N-gram-based text categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval.

Feldman, Ronen. & Sanger, James.(2007). The Text Mining Handbook, Advanced Approach in Analyzing Unstructured Data. Cambridge University Press.

Chiang, Ding-An. & Keh, Huan-Chao. & Huang, Hui-Hua.& Chyr. Derming.(2008). The Chinese text categorization system with association rule and category priority. Expert System with Applications vol. 35. no. 1-2. 102-110.

Khreisat, Laila. (2004) Arabic Text Classification Using N-Gram Frequency Statistics. Tech. report Fairleigh Dickinson University.

Aas, Kjersti. & Eikvil, Line.(1999). Text Categorization: A Survey . Norwegian Computing Center.

Liao, Ciya. & Alpha, Shamim. & Dixon, Paul.(1997). Feature Preparation in Text Categorization, Oracle Corporation.

Mohaqeqi, Morteza. & Soltanpoor, Rreza. (2009). a survey of text mining technique and application. Computer Conference.

Yang, Yiming. & Pedersen, Jan.O. (1997). A comparative study on feature selection in text categorization. machine learning-international workshop then conference.

Mitchell. Tom.(1997). Machine Learning. McGraw-Hill Companies Inc.. USA.

Tan, Songbo.(2006). An effective refinement strategy for KNN text classifier. Expert Systems with Application. vol. 30, no. 2. 290-298.

Elavarasi, Sa. (2011). A Survey on Partition Clustering Algorithms. International Journal of Enterprise Computing and Business Systems,Vol. 1.

Eberhart, Russell. & Kennedy, Jaines. (1995). A new optimizer using particle swarm theory. In: Proceedings of the sixth international symposium on micro machine and human science. 39–43.Saidi, N., Karimkhani, MJ and lively wine, S. (2012). Evaluate the impact of intellectual capital on employee performance across the organization. Rubber Industry Quarterly, 16 (65): 58-66.

Salehi Kerdabad, & Mazandaran, S. (2010). The relationship between innovation and performance. Third National Conference of creativity, TRIZ (TRIZ) and engineering and management of innovation. Tehran: Tehran home teacher.

Schweyer, A. (2004). Talent Management Systems, Best Practices in Solutions for Recruitment, Retention and Workforce Planning, Wiley, New York, NY.

Soltani, A. (2008). Breeding ground for performance management, human resources, policy advocacy scientific journal, 19 (92): 20-23.

Sweem, Susan L. (2009). Leveraging Employee Engagement through a Talent Management Strategy: Optimizing Human Capital through Human Resources and Organization Development Strategy in a Field Study, dissertation, available at: http://www.umi.com/pqdauto.

Tahmasbi, R., H. M., S. & Lyjbary, H. & Khorasani, A. (2013). Developing indicators for identifying talent in strategic business SAPCO (case study: the area of manufacturing and supplying piece), Public Management Research, 6 (20): 97-119.

Tajuddin, M. & Maly Tafti, M. (2008). Why ?, Talent Management magazine Tdbyr.shmarh 191, pp. 62-63.

Taleghani, GR; Amini, S., Ghaffari, A., Dvsy, H. (2013). The relation between talent management and performance of faculty members, Journal of Public Management, 5 (3): 83-102.

True, Muhammad Ali, & Ahmad, I. & R. Seals, H. (2009). Effects of organizational justice on employee performance. Organizational Culture Management, 7 (20): 79-101.

Weiss, A., & mackay, N. (2009). The Talent Advantage: how to attract and retain the best and the brightest. John Wiley & sons, Inc. Hoboken, New Jersey

Yamini, M. (1992). Systems of universities. Tehran: Institute for Research and Planning in Higher Education.