Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 265-278
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/IJNAA.2021.5143

A method for the automatic extraction of keywords in legislative documents using statistical, semantic, and clustering relationships

Jaber Naseri[a,*], Hamid Hassanpour[b], Ali Ghanbari[c]

[a]*Faculty of Computer Engineering, Shahroud University of Technology, Semnan, Iran.*
[b]*Faculty of Computer Engineering, Shahroud University of Technology, Semnan, Iran.*
[c]*University of Science and Technology of Mazandaran, Behshahr, Iran.*

(Communicated by Ehsan Kozegar)

Abstract

Using smart methods for the automatic generation of keywords in legislative documents has attracted the attention of many researchers over the past few decades. With the increasing development of legislative documents and the large volume of unstructured texts, the need for rapid access to these documents has become more significant. Extracting the keywords in legislative documents will accelerate the legislative process and reduce costs. The present study attemptes to extract meaningful keywords from texts by using the thesaurus, which has a structured system to improve the classification of legislative documents. In this method, the semantic relationships in the thesaurus and document clustering were used and the statistical features of different words were calculated to extract keywords. After pre-processing the texts, first the keywords in the text are selected using statistical methods. Then, the phrases derived from the keywords are extracted using semantic terms in the thesaurus. After that, a numerical weight is assigned to each word to determine the relative importance of the words and indicate the effect of the word in relation to the text and compared to other words. Finally, the final keywords are selected using the relationships in the thesaurus and clustering methods. The results of testing various texts from the Parliament of Iran and the Deputy for Presidential Laws indicate the high accuracy of the proposed method in meaningful Keywords extraction.

Keywords: Text mining, keyword extraction, thesaurus, semantic relationships, clustering

*Corresponding author
Email addresses: Naserijaber@Yahoo.com (Jaber Naseri), H.hassanpour@Shahroodut.ac.ir (Hamid Hassanpour), Ali.Ganbari289@Gmail.com (Ali Ghanbari)

Received: April 2021 *Accepted:* May 2021

1. Introduction

The administration of countries is based on accurate and correct laws and the enforcement of correct laws can lead to the progress of the country. Extracting useful knowledge from the high volume of legislative documents can improve the performance of organizations and legislative institutions.

Today, the increasing growth of data and large volume of texts in legislative documents have doubled the need for new methods for the automatic extraction of keywords in texts. The classification of texts using keywords plays a significant role in data recovery. Generating keywords for legislative texts provides more accurate and faster documents for legislation. In these systems, keywords can be generated automatically. The most difficult step in the indexing process is selecting the words which are used for creating index. In this regard, only the words which are candidates for the relevant text are indexed [22].

For this purpose, candidate keywords related to the text is extracted using the knowledge base and legislative documents. This candidate phrase is generated by weighing the different parts in the documents to calculate the semantic relationship between the two texts. Extracting semantic relationships between phrases in various fields of text processing such as checking the similarity between the two texts, correcting the spelling errors, summarizing the text, eliminating the word ambiguity, recovering data, and other fields of natural language processing is of great significance [16].

Due to the increased volume of legislative documents, organizing, analyzing, and displaying them manually is highly difficult and requires a high skill. Extracting keywords manually is a highly difficult and time-consuming process. Thus, it requires an automatic process to access useful data at an acceptable time. The classification of plans and bills in par Liament of Iran is performed manually by experts, leading to human error, inaccuracy, and possibility of mistakes. The prolonged process of reviewing and approving legislative documents increases costs. Access to all of the laws which are relevant or contrary to the plan or bill manually is not possible as it causes a waste of time and reduces the accuracy in approving the laws [20]. A text recovery system should be designed for solving these problems. First, this system should conduct a set of actions to generate an appropriate and efficient index on the words. After creating the system index, it can present the texts related to the words requested by the user [3, 23].

Without keywords extraction, many applications of data recovery such as document classification, summarization, and keyword extraction cannot be defined properly. Thus, an index extraction method was presented in this study to increase the efficiency of text data recovery and classification methods and provide rapid search of data in large sets of documents. Using a structured thesaurus to extract meaningful keywords from texts can improve the classification accuracy of documents.

Accuracy in keywords extraction has the following advantages:

- The automatic extraction of legal documents, publications, and web pages facilitates the reading and searching of data for readers.
- Keywords can be used effectively in organizing data and classifying data.
- The similarity of texts can be recognized rapidly by extracting keywords.

The algorithm presented in this study uses a thesaurus, which is a collection of words, terms, and data related to persion language. The rest of manuscript is organized as follows. The next section reviews the related works. Section 3 addresses the proposed method. Finally, Section 4 presents the results and performance of the proposed method.

2. Review of the literature

In this section, previous studies are reviewed and the findings of them in this field are expressed. Yeton and Modley (2006) proposed an advanced method for extracting keywords based on semantic data to solve the problems in texts [27]. They presented an algorithm based on words indices for documentation which uses machine learning methods and semantic data. Shilpa Dong et al. (2014) declared that almost 80% of the world's data are among the unstructured texts. In order to process the unstructured texts by a computer, a technique is required to extract useful data from unstructured text. In this regard, they suggested that structured data should be first identified and then analyzed to extract valuable data [25]. Eslami Nasab et al. (2014) suggested that one of the most essential pillars in natural language processing and data recovery is to evaluate the similarity of two words or documents. Determining the semantic relationship between two documents, articles, or sentences can be performed through statistical calculations and measuring the similarity between the words. The criteria for the similarity of two words or sentences are used in a wide range of applications such as natural language processing, search query correction, semantic error correction, document comparison, and other applications in data recovery. Furthermore, this study discussed the methods of finding the degree of similarity between documents and words [9]. Rad et al. (2016) presented a method for automatic indexing and keyword extraction for data recovery and text clustering. This study attempted to provide more meaningful keywords from the texts by using linguistic information and thesaurus. Using this method, users can identify the keywords quickly and increase the comprehensiveness of keyword query [19]. Beniwong et al. (2018) reported that text classification is a significant method for organizing, summarizing, and surveying data effectively. This algorithm not only solves the problem before training the model in unsupervised clustering, but also has good effects on the text clustering process. In this study, an algorithm was presented for textual data based on deep learning representation. The parameters of the general framework of this model were defined in such a way that most of the training in the deep classification model formed the model using the labeled data set in the domain. Second, the domain detector was added to the model. Finally, the domain-compatible parameters were used as a model [5].

3. The proposed method

In this study, automatic keyword extraction was conducted using statistical and semantic methods. First, the word frequency in the document and the word frequency in all documents are calculated and the number of documents is obtained. Then, repetitive words and probable words are calculated according to the simultaneous occurrence. Finally, the final keywords are extracted. Further, semantic connection, any kind of dependence, and conceptual relationship between two words, phrases, or texts are recognized in this study. A thesaurus with a structured system was used in this method to extract more meaningful keywords from the texts and improve the classification of Persian texts. At this step, the relationships between general and specific words are identified and the accuracy of text classification is improved after processing the text using the thesaurus.

This study attempted to provide an appropriate method for automatic index construction in legislative texts using the relationships between words thesaurus. In addition, this study provided a solution for weighing words using the tf-idf algorithm (Term frequency inverse document frequency). Here are the concepts of basic definitions and the steps of the proposed method.

3.1. Legislative documents

Legislative documents include the plans and bills discussed in the Parliament of Iran. In addition, the texts of these plans and bills are compared and analyzed with the current laws and sent to the

parliament after extracting the items which are relevant or contrary to the laws of Iran.

3.2. Text classification

If we have a set of texts $D = \{(d_1, y_1), \dots (d_i, y_i), \dots (d_n, y_n)\}$ so that n represents the number of texts, $di = \{w_{i,1}, \dots, w_{i,k}, \dots, w_i, |di|\}$ represents the i-th text of this set, and $w_{i,k}$ is the word k in the i-th text, y_i indicates the class to which the text belongs, i.e. $y_i \in C$ so that $C = \{c_1, c_2, \dots, (c|c|)\}$ indicates the set of predefined classes in the system. The goal of text classification is the inference of a function f so that we can have $y_i = f(d_i)$. In this regard, text classification is a Boolean value for each pair of $(d_j, c_i) \in D \times C$ where D represents a set of texts and C indicates a set of predefined classes. T value (true) indicates that text d_j belongs to class c_i, and F value (False) shows that text d_j does not belong to c_i.

3.3. Thesaurus

A thesaurus is a set of words, terms, and information related to a specific field of language. In addition, a thesaurus refers to the words which crystallize information and includes the meanings of the intended text and helps extract keywords from the text, recovers text data, and helps in text indexing [15]. Synonyms and dependent words whether general or specific are identified using the thesaurus after the main features extracted from the text preprocessing section. In other words, synonyms and antonyms whether general and specific are extracted and saved somewhere for each main word of the text and these words are later used for weighing.

The purpose is to select a word from synonyms or general and specific words in the text instead of considering a weight for each word and assigning a specific weight coefficient[15].

3.4. Semantic relationships

Semantic relationship is a level higher than a variety of relationships in the network of words. In fact, all of the relationships expressed in words and phrases indicate a kind of semantic relationship. In general, two phrases may be semantically related to each other but have no place in vocabulary structures. For example, there is a semantic relationship between “night” and “moon”. In fact, two words can have a semantic relationship with each other if they are used with each other, or one of them can affect the other one, or they can be associated with each other in speech and thoughts. Since the similarity between the two texts indicates a kind of relationship, the vector space model method can be considered as one of the methods for calculating the semantic relationship. The main hypothesis is that the two related texts have similar words. The latent meaning analysis (LSA) method can be used for calculating the semantic relationship, in which the phrases are transferred to a new meaning space by mapping. The main idea in this method is to reduce the dimensions using singular value analysis [13].

3.5. Extracting keywords from Persian texts using the statistical method

After the normalization process, deleting static words, fragmenting words into text, and rooting out words are among the most critical steps in text processing for keyword extraction. Keyword extraction provides the automatic detection process of the terms used in a document.

This process is normally conducted in three steps. First, a set of words or phrases are selected as a candidate. Then, the attributes which make the word or phrase as keywords are calculated for each candidate. Finally, all of the candidates are extracted by combining the attributes in a formula and using statistical methods such as tf-idf (this method is explained below) [11].

3.6. Calculating the number of keywords in text documents

Extracting the keywords from documents is one of the most significant prerequisites in the process of clustering, classifying, and extracting the required data. The appropriate number of key phrases for each text has a logarithmic relationship with the number of keywords in that text.

The method of calculating the number of keywords is extracted from the following mathematical equations for each document.

$$\text{Number of keywords for each text} = 2.21*\log(\text{WN})-3.43$$

where WN represents the total keywords [11, 13].

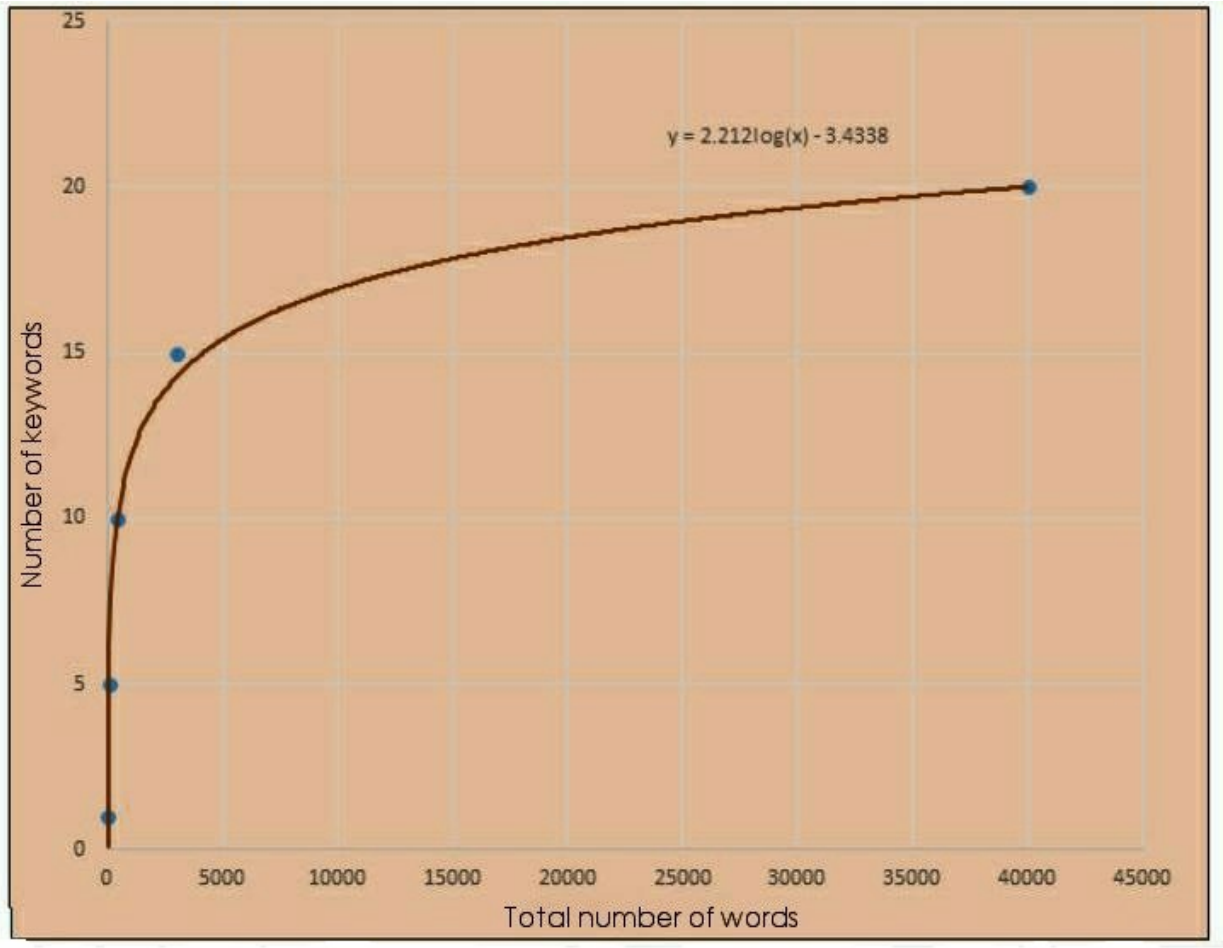

Figure 1: Number of keywords in terms of the total number of words in each document [11]

As shown figure 1, if the total number of words exceeds 10000, there will be a constant and uniform slope for calculating the number of keywords, ranging between 15 and 20.

In this step, general words are deleted in the text and then the words are rooted and the generated general words are deleted again. The next step is to create a dictionary which is a collection of words that covers all the words in the documents. In other words, any word that has appeared at least once in a set of documents is included in the dictionary [10].

After pre-processing the text, the keywords are extracted using the TF-IDF method.

3.7. TF-IDF method

The TF (Term frequency) method is a widely used method for extracting keywords and its steps are as follows. First, a vector of the words within a document or text is created. Assigning points to each word based on the repetition of that word is performed in this step. Then, the scores are sorted in a descending order. Finally, the selection of the number of keywords with more scores is applied. In this way, we consider the repetition rate of a word in a document versus the number of repetitions in the word set of documents and determine the weight of words using term frequency and inverse document frequency. Equation (1) is related to term frequency criterion, Equation (2) is related to the inverse document frequency criterion, and Equation (3) is related to the calculation of the weight of each word [32].

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{max\{f(w,d) : w \in d\}} \qquad (1)$$

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \qquad (2)$$

where $f(t,d)$ represents the number of repetitions of the word t in the document d and $maxf(w,d)$ indicates the number of most repeated words in the document d in which N shows the total number of documents and the denominator indicates the number of documents in which there is the word t. Finally, the weight of each word is calculated based on Equation (3) as follows.

$$tf - idf(t,d,D) = tf(t,d) \times idf(t,D) \qquad (3)$$

3.8. Clustering using the K-means algorithm

The K-means algorithm is performed as follows:

1. Set A is divided into K clusters $[\prod 1, \ldots, \prod n]$ and the following relation is established among the clusters. This relation is shown in $\sum_{i=1}^{k} \prod i = A$.

   This division can be conducted by humans or by machine. Clusters can be shared with each other and can be together without sharing.

2. Central vector $\overrightarrow{C_i}$ is calculated for each cluster.

   Creating the central vector $\overrightarrow{C_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} az \qquad (4)$

   where n_i represents the number of documents in cluster $\prod i$ and az shows the vector of the z-th document in cluster $\prod i$.

   Normalizing the central vector $\overrightarrow{C_i} = \frac{\overrightarrow{C_i}}{\|Ci\|}; i = 1, 2, 3, \ldots, k \qquad (5)$

   The new document a_j is placed in a cluster while being added to the set of documents whcih has the maximum similarity to the central vector of that cluster. In other words, $a_j \in \prod m$ if $a_j\overrightarrow{Cm} > aj\overrightarrow{C_i}$ for each C_i and $(C_i, \ldots, C_m)$ are vectors.

3. General matrix C which includes all central vectors is created as follows.

$$C = [\overrightarrow{c_1}, \overrightarrow{c_2}, \overrightarrow{c_3}, \ldots, \overrightarrow{c_k}] \qquad (6)$$

While recovering with query q, the cluster(s) which have the maximum conformity with q are searched. This amount of conformity is obtained by calculating the cosine of two vectors $\overrightarrow{q^T}$ and $\overrightarrow{C_l}$ as shown in follows.

$$\overrightarrow{q^T}\overrightarrow{C} = [\overrightarrow{q^T}\overrightarrow{c_1}, \overrightarrow{q^T}\overrightarrow{c_2}, \overrightarrow{q^T}\overrightarrow{c_3}, \ldots, \overrightarrow{q^T}\overrightarrow{c_k}] \qquad (7)$$

.

Cosine of vector $\overrightarrow{q^T}$ with cluster $\prod i$ is more than the cosine of vector q^T with cluster $\prod j$, then $\prod i$ has more conformity with q than cluster $\prod j$ [31, 29].

3.8.1. Text clustering using incremental algorithm in detecting the key subjects of text

Using the text clustering method and incremental algorithm, news texts and legislative documents with the same content concepts can be automatically discovered and organized and the accuracy and efficiency of clustering can be optimized. This algorithm reduces the real-time performance of the process and the reduction dimension in the text feature vector. In addition, it calculates text similarity and text processing and simplifies the scale of text similarity complexity. The steps of data processing in texts with this algorithm are such that text vectors are created and clustering analysis is applied for text vectors after reducing the dimensions on the text [28]. Fig. 2 displays incremental text clustering based on similarities.

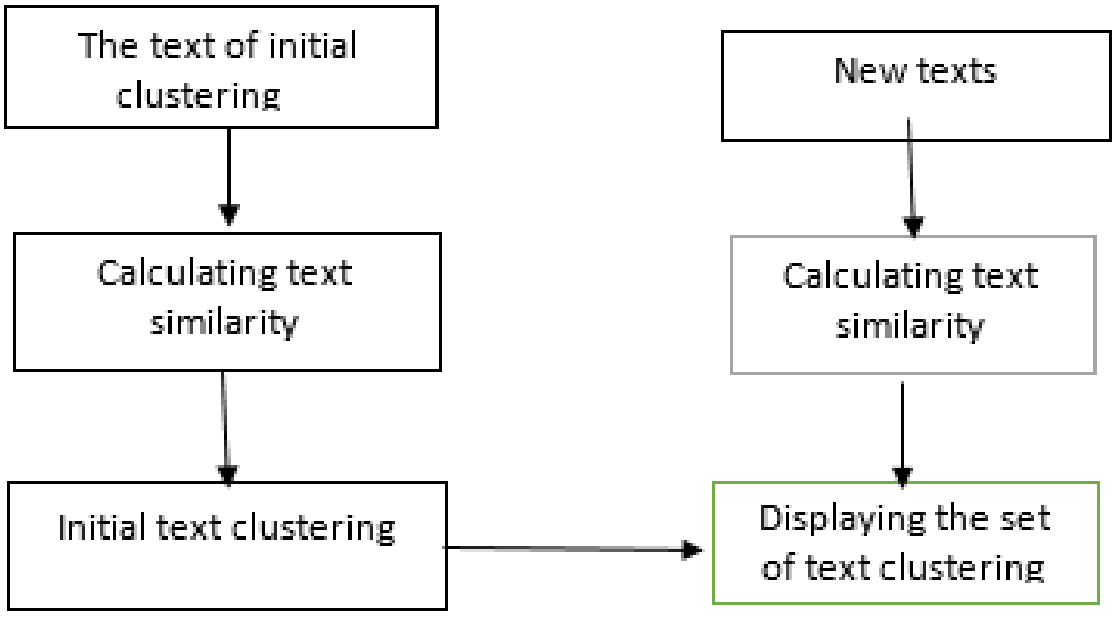

Figure 2: Incremental text clustering based on similarities

4. Steps of the proposed method

4.1. Text pre- processing and processing

After normalizing the documents, the redundant words are removed, the words in the text are fragmented, and the words are rooted. Then, conversational phrases are converted into formal, spelling errors are checked, and semantic and statistical similarity is calculated [6, 14].

4.2. Automatic keyword extraction

After pre-processing the text, first the keywords are selected and the weighing of these words is conducted by tf-idf method [30]. Then, the keywords are clustered and the number of clusters is determined based on the keywords.

4.3. Text clustering

K-means clustering algorithm for large datasets clusters the data dynamically so that it first calculates the threshold value as K-means centers and the number of clusters is formed based on this value. In each K-means iteration, if the Euclidean distance between two points is less than or equal to the threshold value, the two data points are placed in the same group. Otherwise, it creates a new cluster with a different data point [17, 12].

While extracting the keywords, we cluster the keywords, and determine the number of clusters based on the keyword set.

5. Evaluating the proposed method

To implement the proposed method after classifying the texts, which includes (text processing and selecting text features and weighting the words), then determine the weight of the words in the text and then find the keywords in the text. For this purpose, we do the following:

1. examine the words in the text and find the number of repetitions of the word in the text
2. We calculate the weight of words using two criteria term frequency and inverse document frequency

Thus, the texts with the similar subject are expected to be in a cluster. This method randomly selects some points as the number of clusters required by the text and then the data are attributed to one of the clusters based on the degree of proximity and similarity, resulting in creating new clusters. By repeating the same procedure, new centers are calculated by averaging the data and the data can be attributed to new clusters. This process continues until the data are unchanged. In this method, a feature vector is used for displaying the above text in clustering, which includes keywords and the repetition rate of the words in the text. In order to cluster the new text, first the text feature vector is created and then compared with the cluster feature vector. If the new cluster is detected, it will be added to the list of clusters, otherwise; the growth of clusters will stop. In this way, the text is shown by a word matrix so that each entry displays the number of repetitions of each word in the text [18, 4].Fig. 3 presents the proposed model design in three phases.

In the first phase, ontology is conducted on the texts [24]. In this method, the quality of data recovery in texts increases and the relationship is provided between semantic concepts in texts. Thus, the required structure is extracted in the texts using text processing techniques. In the second phase, data structure is extracted and text is pre-processed. In addition, keyword generation, refinement of changes and sensory vocabulary (words that are related and interdependent) are made and determination of relationships between features are conducted. Using key features and phrases in this phase, the relationships are conducted in a specific format. Finally, the phrases in the text and their characteristics are grouped and the text is converted into semantic data. As a result, the text keywords are extracted.

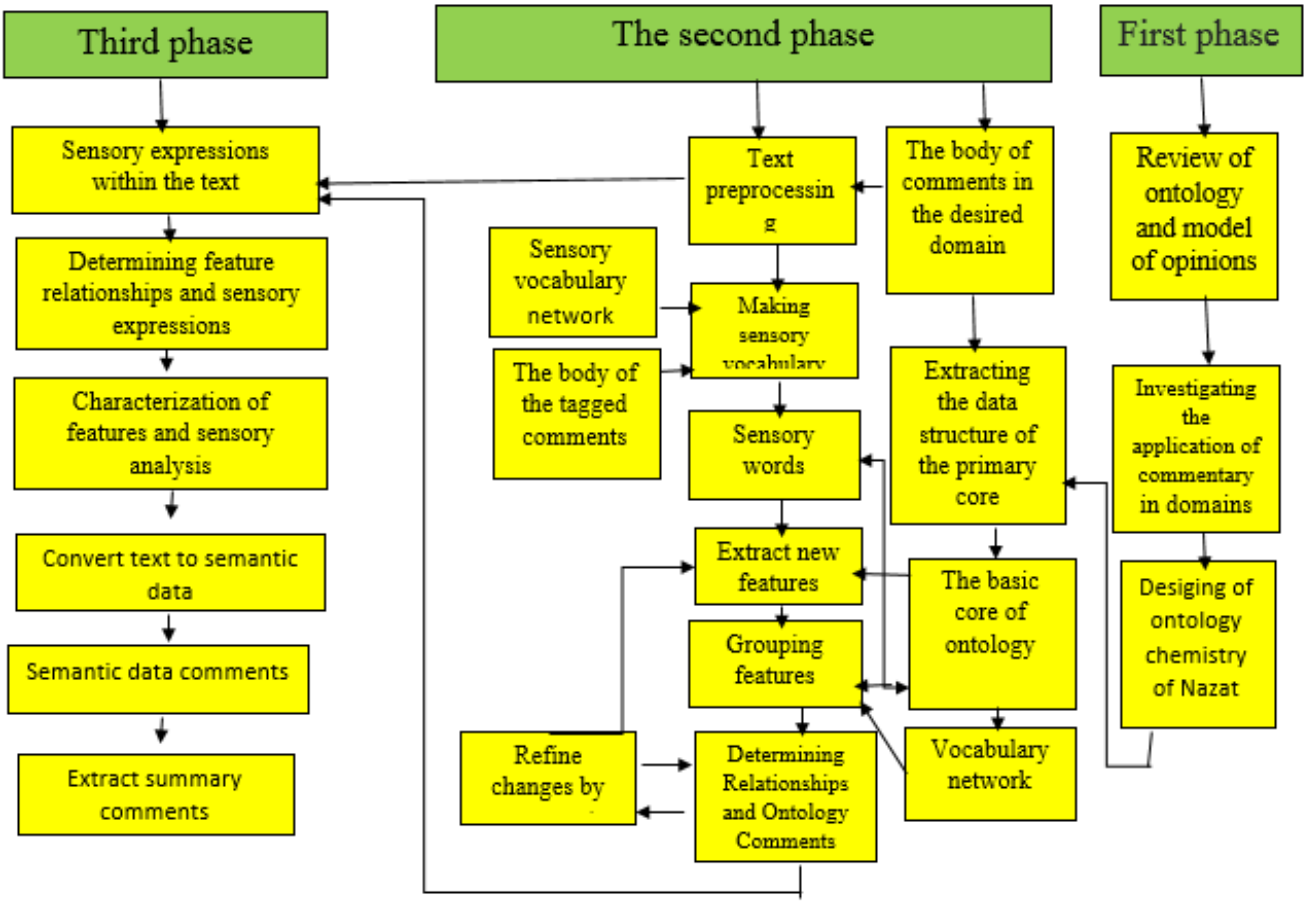

Figure 3: Designing the proposed model

Table 1: General information

Row	Thematic category	Number of documents	Average number of words in documents
1	Cultural	150	170
2	Economic	130	177
3	Agriculture	120	127
4	Healthcare	80	140
5	Sports	60	95

6. Selecting legislative documents for testing

To test the proposed method, a set of legislative documents was collected in five different categories from the database of the Parliament of Iran [7, 8].

The set of legislative documents is selected to cover a wide range of words related to each domain. In this case, the ability of the proposed method to detect texts with different words in a specific domain is demonstrated. Here are the implementation steps. In pre-processing step, the extra words of each text should be removed and the root of the extracted words and the keywords of the text should be entered into the text database. For this purpose, a file containing common additional words is considered and the desired text is read word by word and compared to each other. If the desired

word is in the additional words file, it is removed from the text; otherwise; the record of the desired text in the database is added. At this stage, some common prefixes and suffixes are separated from the remaining words and the root of the words remains. Tables 2 and 3 for each thematic category indicate the average volume of words removed in the word preprocessing operation and the percentage presented to the user as keywords.

Table 2: Processed files

Row	Thematic category Documents (plans and bills)	Average number of words in documents (Plans and bills)	Average number of words after removal and rooting
1	Cultural	123	80
2	Economic	199	135
3	Sports	204	149
4	Social	150	100

Table 3: Results of processed files

Row	Thematic classification of documents (plans and bills)	Percentage of extra words which were incorrectly detected as part of the main word.	Percentage of main words which were incorrectly detected as part of the main word.
1	Cultural	22%	12%
2	Economic	18%	11%
3	Sports	20%	9%
4	Social	16%	10%

6.1. Finding the number of keywords in text and weighing the main words

In this step, after performing the text preprocessing step, during which the main words of the text are extracted. It is time to count the number of repetitions of the words in the text, and then we calculate the weight of the words and identify the family of texts using a word dictionary. And for all the words we consider a candidate in the text and a weight coefficient is added to the frequency of repetition of the word to the candidate in the text and then we get the appropriate weight between the words based on the semantic relationship. Thus, based on weights, the most appropriate keywords are extracted in the texts [21]. Using the algorithm presented in [26] by Mr. George, the appropriate weight is obtained according to the semantic relationship between the words in the text, upon which the most appropriate key indicators of the text are extracted. Using this algorithm which aims to find the degree of repetition of synonyms in the text, the desired thesaurus is checked for each main word. If read word is synonym to another word which was previously read in the text, a weight percentage equals to β coefficient which is between zero and one is added to the candidate word counter instead of adding a repetition of the new read word. Here, a weight percent β is placed instead of the words repetition and synonyms.

By seeing the specific and general words, a certain weight percentage as α is added to the weight of the candidate word. In order to manage such properties, first the feature vector should be obtained and then applied on keywords. Then, we spread the keywords on other features and this method has a high accuracy and less time overhead.

Here, it can be logically guessed that the value considered for the coefficient of synonymy is more than the coefficient obtained for general and specific words because synonyms come from the same root and these words have a more similar subject than general and specific words. Based on this algorithm, many experiments and comparisons are conducted to obtain the values of α and β in an optimal and appropriate way and we conclude that if $\alpha = 0.2$ an d $\beta = 0.4$, it would be a good choice [21]. The results are reported in Table 4 while the comparison of the accuracy of the proposed method with other methods is shown in Table 5.

Table 4: Results of the proposed method

The proposed method	Using the thesaurus using the method $\alpha = 0$, $\beta = 1$	Using the thesaurus using the method $\alpha = 0.25$, $\beta = 0.3$
Accuracy	73/21	73/68

Table 5: Comparison of the proposed method with other methods

Accuracy (percentage)	Methods
68/73	The proposed method
64	Arasteh's method [21]
67	Aghalebandan's method [1]
63	Parliament database method

In the Parliament database, the criterion of comparison based on the plans and bills which have been previously referred to the Parliment and reviewed in the commissions are evaluated. In this comparison, the work of the parliament is performed experimentally and manually based on experiences of the experts in the deputy of laws in reviewing the plans and bills. However, since it is conducted with experienced and skilled experts, the accuracy of extracting the keywords of the texts reaches 63%.

In the method of Alagheband et al., a combined method of clustering based on the center of clusters, K-means, along with the SVD method was studied in which clustering was performed by introducing two new concepts of neighborhood points and relationships [1].

Using SVD can reduce the noise effect by preserving the underlying structure in the feature space. Using the presented concepts, a method was introduced for selecting the primary centers of the clusters and a relationship for calculating the similarity of the text vector with these concepts.

Finally, different experiments were performed on a set of 56 texts. In this experiment, the K-means algorithm was used and neighborhood and link concepts were applied to improve the classification of texts to 67% using the SVD algorithm [1]. In the method of Arasteh et al., the semantic clustering method of Persian words was conducted on a collection of Persian texts and articles. In this experiment, the K-means algorithm was used to improve the clustering and classification of Persian texts to 64%. [2]

7. Evaluation criteria of the proposed model

In order to evaluate the proposed model, the data of the Islamic Consultative Assembly and the data of the Vice President for Compilation and Revision have been used. These data are kept in the Law Office of the Islamic Consultative Assembly. And the output of our work and the correctness of our performance are compared and evaluated based on the plans and bills that have already been reviewed and approved by the Islamic Consultative Assembly.

8. Conclusion

Today, with the increasing volume of information, the existence of intelligent systems for automatic classification of texts is essential. In this article, an optimal method for extracting keywords and indexing Persian texts was presented. This method can be used in the process of thesaurus, weighting, linking educational semantic data processing to extract the best qualities of candidates for each word. In this method, after pre-processing the legislative documents, first the index words in the texts were selected using statistical and semantic methods. Then a key in the text is extracted using semantic terms in the thesaurus and then to determine the relative importance of the words a numerical weight is assigned to each word which indicates the effect of the word in relation to the subject of the text and in comparison with other words used in the text. Finally, the final keywords were selected using the relationships in the thesaurus and clustering methods.

According to the results obtained from the analysis and outputs of this article, it seems that for words and phrases with general relations, specific weight coefficient 0.2 also for the words of the same family, a weight coefficient of 0.4 should be considered. Experiments with data and experimental texts indicate the proper selection of these parameters in the recognition of keywords in the texts. Of course, the results obtained from the algorithm, which expresses the weighting factor, are also logically acceptable, because words of the same family that have a common root are generally in the same subject and are used in a special category and therefore will have a higher coefficient.Also, the proposed method is fully complied with, evaluated and tested with the criteria of the Islamic Consultative Assembly database.

References

[1] R. Alagheband and MR. Saeedi Mohammadi, C*lustering of texts based on category center using svd method and using neighborhood points*, International Conference on Persian Line and Language Processing, (2012).

[2] A. Arasteh, M. Elahimannesh and A. sharif, B.minaei-Bidogli, *Semantically clustering of persian words*, Conference on Persian language proseshng semnan, Iran, (2012).

[3] M. Beh Shameh and H. Bashiri, *From clustering and summarizing documents for latent semantic indexing*, Hamedan University of Technology, (2009).

[4] A. Benabdellah and A. BenghabritJamane bouhaddou, *A survey of clustering algorithms for an industrial context*, Second International Conference on Intelligent Computing in Data Sciences (ICDS). (2018).

[5] W. Binyu , L. Wenfen, L. Zijie, H. Xuexian , W. Jianghong and L. Chun, *Text clustering algorithm based on deep representation learning* , The 2nd Asian Conference on Artificial Intelligence Technology, (2018).

[6] C. Charul and I. Aggarwa, *A surves of text clustring algorithms*, Watson research center yorktown heights, NY. Cheng Xiang Zhai university of Illinois at Urbana-Champaign Urbana. (2012).

[7] Database of the Deputy of Laws in Parliament of Iran.

[8] Database of the Vice President for Codification and Revision of Presidential Laws.

[9] M. Eslami Nasab and R. Javidan, *A method to find the semantic similarity of articles and documents*, Shiraz university of technology, (2014).

[10] H. HudhudKian, *Assessing the similarity of documents on the web*, Master thesis, Shiraz university, (2011).

[11] Iran big data house. Extracting keywords from Persian text.

[12] Yi. Junkai, Zh. Yacong, Zh. Xianghui, and W. Jing, *A Novel text clustring approach using deep learning vocabulary network* , college of information science and technology, beijing university of chemical technology, Beijing , China. (2017).

[13] H. Kamyar, *New method of semantic weighting of words in word processing applications*, Master Thesis, Ferdowsi University of Mashhad, (2011).

[14] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, and R. Nisbet, Elsevier, the Seven Practice Areas of Text Analytics. (2012).

[15] M. Mohammadi, M. Ataloui, *Keyword extraction of Persian documents*, Sharif University of Technology, (2007).

[16] A. Mosallanejad, J. Davoodi Moghadam and A. Ahmadi, *Presenting an efficient algorithm for extracting semantic relationships in documents based on Wikipedia knowledge base*, 23rd Iranian Conference on Electrical Engineering, Sharif University of Technology, (2015).

[17] P. Nerurkara, A. Shirkeb, M. Chandanec and S. Bhirudd, *Empirical Analysis of Data Clustering Algorithms* , 6th International Conference on Smart Computing and Communications, ICSCC, Kurukshetra, India, (2017).

[18] P. Pantel and D. Ravichandran, *Automatically labeling semantic classes*, In Proceedings of HLT-NAACL. (2004).

[19] F. Rad, H. Parvin and A. Dehbashi , *Introducing a new method for automatic indexing and keyword extraction for information retrieval and text clustering*, (2016).

[20] *Research Center of the Parliament of Iran* , Artificial intelligence and legislation, (2018).

[21] V. Rezaei, M. Mohammadpour, H. Parvin and S. Nejatian, *Providing a method for extracting keywords and weight of words to improve the classification of Persian texts*, (2017).

[22] Sh. Safari, *Automatic production of keywords for scientific documents using semantic relations*, Master Thesis, (2015).

[23] F. Sedghi, H. Bent Al-Huda and A. Turkaman Rahmani, *Using a set of finders to classify documents (an approach based on artificial security)*, Iran University of Science and Technology, (2014).

[24] M. Shams Fard , A. Abdullahzadeh, *Extraction of conceptual knowledge from text using linguistic and semantic patterns*, Amir Kabir University of Technology, (2002).

[25] D. Shilpa Dang and H. Peerzada, *Text Mining: Techniques and its Application*, November (2014).

[26] G. Tasatsaronis, I. varlamis and M. vazirgianise ,*Text relatedness based on a word thesaurus"Jornal of artificalligence*, 37 (2010) 1-39.

[27] W. Witten and H. Medelly, *Thesarus based automatic key phrase*, Conference on digital, (2006).

[28] Z. Xiaoming, L. Zhang , *Automatic topic detection with an hncermental clustering algorithm*, Beihang University .Beijing china .Lncs 6318, (2010) 344.

[29] L. Yuri , R. Anand, D. Jeffrey, *Translated by Mehdi Esmaili*, Kavoshdadegan, (2017).

[30] M. Zakir Hossain, M. Nasim Akhtar, R.B. Ahmad and M. Rahman, A *Dynamic k-means clustering for data mining*, Indonesian journal (2019).

[31] M. Zakir Hossain, M. Nasim Akhtar, R.B. Ahmad and R. Mostafijur, *A dynamic K-means clustering for data mining*, Indonesian Journal of Electrical Engineering and Computer Science, 13 (2) (2019) 521-526.

[32] Zh. Zhiling, J. Zhu, D. Liang, H. Li and L. Yu.Guoql, *Hot topic detection based on a refined TF-IDF algorithm*, National Natural Science Foundation of China, (2016).