Web mining based on word-centric search with clustering approach using MLP-PSO hybrid

Reza Samizadeh* and Samaneh Tafahomi

Department of Industrial Engineering,
Alzahra University,
Tehran, Iran
Email: rsamizadeh@alzahra.ac.ir
Email: s.tafahomi25@gmail.com
*Corresponding author

Abstract: With web development, sometimes in keeping track of information on the web, the semantic meaning of words is not important, and the mere presence of words in the text is enough to extract information. In this research, the word-centric search method is presented to prepare web data for clustering. Multi-layer perceptron networks are one of the most successful neural networks for learning, clustering and prediction. The researcher clusters the web data from the word-centric search method by using the K-means method and considers the results of clustering as the expected output of the MLP neural network. Considering that the weights of the neural network are selected randomly, it may not be in the best amount after the network training. Therefore, by using an optimisation algorithm for particle swarm, its effect on performance of the final neural network has been investigated in the training and initial weighing step.

Keywords: web mining; clustering; multi-layer perceptron neural networks; particle swarm optimisation algorithm.

Reference to this paper should be made as follows: Samizadeh, R. and Tafahomi, S. (2022) 'Web mining based on word-centric search with clustering approach using MLP-PSO hybrid', *Int. J. Business Intelligence and Data Mining*, Vol. 20, No. 1, pp.35–55.

Biographical notes: Reza Samizadeh is an Assistant Professor of the Alzahra University, where he teaches systems analysis, customer relationship management, strategic planning, enterprise resource planning, design of information systems and simulation. His research interests focus on CRM, digital transformation, supply chain, banking, data mining and supply chain. He received his PhD in Industrial Engineering (IT) from Tarbiat Modaress University, MS in Systems Engineering from Sharif University of Technology, and BS in Industrial Engineering (Industrial Manufacturing) from Sharif University of Technology.

Samaneh Tafahomi obtained his Master's in Industrial Engineering from the Alzahra University, Iran in 2018. Her research interests focus on big data, data mining and web mining. The topic of her thesis is 'Web mining based on word-centric search with clustering approach using MLP-PSO hybrid'.

1 Introduction

Web mining refers to searching and exploring in the vast World Wide Web for the purpose of finding specific data and information. Website is a branch of data mining and data mining in the term knowledge is called discovery database. Web mining to search for desired data should be performed in a way that the best results are presented to the user in the shortest time possible. With the rapid growth in data in the World Wide Web, there is a need to improve the fast and secure ways to perform web-based operations. Due to the large amounts of information available in the World Wide Web, users always expect to see the best results they are looking for at the beginning of the search results list. Therefore, providing a method to improve the results presented to the user is the main motive for this research. Sometimes, in keeping track of information on the web, the semantic meaning of words is not important, and the mere presence of words in the text is enough to extract the data. In this case, only the word itself is considered not the meaning of it. In this way, the web is searched based on keyword. In fact, by using clustering techniques, information and texts on the web can be classified for evaluating and producing knowledge. Hitherto, different similarity measurement metrics have been introduced to measure web clustering data. In 2014, Lin et al. introduced SMTP similarity measurement metrics. In this criterion, the similarity between the two documents is calculated based on the existence or non-existence of the feature. In this study, the use of selected keywords has been used as a feature. As we use the neural network for predicting and clustering, we cluster. This study concludes that evolutionary algorithms affect the performance of the neural network.

1.1 Research background

The term ‘web mining’ was first introduced in 1996 by Etzioni in an article titled ‘The World Wide Web: quagmire or gold mine’. In this paper, web mining is described as a task-oriented method (Etzioni, 1996). In 1997, Cooley et al. from a data-driven point of view, addressed the full definition of web mining in an article entitled ‘Web mining: information and pattern discovery in the World Wide Web’. The first specialised panel discussion in this area was held at the 9th International Conference on Electrical and Electronics Engineers in the field of artificial intelligence entitled ‘Web mining illusion or reality’ (Srivastava and Mobasher, 1997a, 1997b). Each year, various workshops on web mining have been conducted since 1999 by the Association of Computer Machines (https://www.acm.org/) and since 2001 by the Society for Industrial and Applied Mathematics (https://www.siam.org/).

Web mining methods may be divided into three types of web content exploration, web structure exploration, and web user exploration, based on what kind of data they are exploring (Jokar et al., 2016).

Extracting content from the web pages has been done by using topics such as text, images, multimedia, and so on. Exploring the web structure is associated with the linking structure of web pages and is used to extract information from these structures. Exploring the web user is related to user’s behaviour. Recorded user activities data in weblogs are used to extract important information (Kosala and Blockeel, 2000). In some texts, web mining is referred to as the exploration of the web users file, which is called the fourth method (Fürnkranz, 2005; Wang et al., 2005; Borzemski, 2006).

Explore the web users file to discover the file based on their behaviour on the web which is most commercially used in identifying customers.

Integrating the web content exploration and applied web exploration is also possible. Hyperlink recommendations solve the problem of quick and easy access to information on web systems. Kazienko and Kiewra (2003) proposed an integrated approach to web applied exploration and web content exploration. Interesting documents are potentially active for the user based on usage patterns. The term of automatic selection and the distinction between using the web regarding the time of visit was introduced to estimate the effectiveness of the method (Kazienko and Kiewra, 2003).

Taherizadeh and Moghadam (2009) extracted the text content of web pages through a duplicate word sequence. The text content was combined with the web server logs to study the rules of the user behavioural association. The result of the proposed system helps to better recommend web personalisation, web construction and web user profiles.

The relationship between the web content exploration and the web structure exploration has been discussed in Gedov et al. (2004) research. In that research, the content of webpage is compared with the information defined by the website's structure. Each web page is described with a set of keywords. This information is combined with a structure link that generates content based on the description. This comparison helps to access the semantic information of the web page and its adjacency (Gedov et al., 2004).

Chou et al. (2010) proposed a method based on customer online conduct behaviours by analysing their guidance patterns through web mining and artificial neural network manufacturing, which helps them to search more effectively in a web-based environment and used to predict the potential customers need in the future.

Jain (2010) presented a clustering summary; a summary of known clustering methods, discussing major challenges and key issues in the design of clustering algorithms, and pointing to some of new and useful research paths, including semi-monitoring clustering, group clustering, the choice of synchronous feature during clustering of data and large-scale clustering of data.

Text mining is a method for identifying relationships and patterns and methods in textual data. Rezaeian et al. (2017) have used clustering of key words using the K-means algorithm.

Measuring the similarity between documents is an important practice in the field of text processing. In a study by Lin et al. (2014), a new similarity criterion, the SMTP model is proposed. To calculate the similarity between two documents according to the feature, the proposed criterion considers the following three conditions:

a the attribute is seen in both documents

b the attribute is only visible in one document

c Feature is not visible in any of the documents.

For the first condition, the similarity increases with decreasing the amount of difference between the two features involved. In addition, the amount of difference is usually measurable. For the second condition, a constant value for similarity is considered. For the last condition, the attribute have no share in similarity. The proposed criterion has expanded to measure the similarity between two sets of documents. The effectiveness of the criteria has been evaluated in several actual datasets for clustering and text categorisation issues. The results show that the performance of the proposed criterion is

better than the results obtained by other criteria (Lin et al., 2014). Orhan et al. (2011) introduced a perceptron-based neural network classification model as a diagnostic decision support mechanism in the study of epilepsy treatment. Probabilistic distributions are calculated according to the distribution of wavelet coefficients in the clusters and then used as inputs of the perceptron neural network model (Orhan et al., 2011).

Zhang et al. (2007) proposed a hybrid algorithm for combining particle swarm optimisation (PSO) algorithm with back propagation algorithm (BP), called the PSO-BP algorithm, for training neural network weights (FNN)[1]. Hybrid algorithm cannot only use the powerful global PSOA search capability, but also has the ability to search local and strong BP algorithms. In this paper, a new automatic weight selection strategy for the PSO algorithm is introduced. In the PSO-BP proposed algorithm, they adopted an evolutionary technique to transfer the search for particle swarm to the search for gradient descending. Experimental results show that the proposed PSO-BP hybrid algorithm is better than particle optimisation algorithm (APSOA)[2] and BP algorithm at convergent speed and accuracy (Zhang et al., 2007).

Many efforts have been made to predict the infinite compressive resistance of rocks using an artificial neural network post-release that suffers from disadvantages such as the slow pace of learning. Momeni et al. (2015) proposed an ANN hybrid model based on PSO to predict the infinite compressive resistance of rocks. In their research, predicting the performance of the proposed ANN-PSO hybrid model has been studied in comparison with the ANN model based on the comparison between the R2 determination coefficients. The results show that, in predicting the infinite compressive resistance of rocks, the proposed ANN-PSO hybrid model is superior to the ANN model.

Za'in et al. (2017) proposed an evolving web news mining framework based on evolving type-2 classifier (eT2Class). The eT2Class adopts an open structure that can be used in non-stationary environments and works on a single pass learning mode that is applicable for online real-time applications (Za'in et al., 2017).

2 Research methodology

This section introduces the concepts and tools used in this research. First, the concepts and applications of artificial neural networks are outlined, and then concepts of evolutionary algorithm and PSO algorithm are investigated. Then the source data used in this research is introduced and explained.

2.1 Web mining

Web mining refers to all data mining activities and related techniques that are used to automatically discover and extract data from web documents and web services. Behind the vast resources of information on websites, the structural information and data available on the web and its servers, the knowledge lies deep beneath that is difficult to access normally and is very useful for users, managers and administrators of the web environment. Web mining and related techniques allow discovering and identifying this hidden knowledge and information. Depending on the types of function, web mining involves three methods: explore web content, explore the web structure, explore the web practically. Web mining involves four main stages:

1 *finding source:* this step involves retrieving the requested web documents

2 *information selection and preprocessing:* at this step, specific information from the retrieved documents is automatically selected and pre-processed

3 *generalisation:* at this step, generic patterns are automatically discovered on one or more websites

4 *analysis:* at this step, the patterns obtained at the previous step are validated and interpreted.

In the first step, data is retrieved from sources on the web, such as electronic newsletters, newsgroups, HTML documents, text databases, and so on. The selecting and preprocessing stage involves any process of converting the retrieved data in the previous step. This preprocessing can include reducing words to their root, eliminating extra words, finding expressions in the text, and converting the representation of data into a relational form or first-order logic. In the third step, data mining and machine learning techniques are used for generalisation. It should also be noted that users play an important role in the process of extracting information and data from the web. This point is especially important in the fourth phase.

In this way, web mining is the process of discovering information and unknown and useful knowledge from the web. This process implicitly includes discovery in databases (KDDs)[3]. In fact, web mining is a developed species of KDD that acts on web data.

2.2 K-means clustering algorithm

Organising data into meaningful groupings is one of the most basic ways of understanding and learning. The Cummins K-means method is one of the data clustering methods in data mining. In the K-means algorithm, the *k*-member (*k* is the number of clusters) is randomly selected among the *n* members as cluster centres. Then the *n-k* remaining members are assigned to the nearest cluster. After assigning all members, the cluster centres are recalculated and assigned to the clusters according to the new centres, and this continues until the centres of the clusters stay fixed.

The best clustering is to maximise the total similarity between the cluster centre and all cluster members, and minimise the total similarity between cluster centres. In text mining for clustering the text, using the K-means algorithm based on the keyword, we can cluster the information (Kazienko and Kiewra, 2003).

2.2.1 Similarity measurement methods used in clustering

In this research, two criteria of SMTP model and Euclidean distance are used for clustering.

SMTP model

The SMTP model is to calculate the similarity between two documents according to the feature. The proposed criterion considers the following three conditions:

a the attribute is seen in both documents

b the attribute is only visible in one document

c the attribute is not seen in any of the documents.

For the first case, a lower limit of 0.5 is considered, and by reducing the difference between the numbers of attributes in the two documents, the similarity is increased, which is calculated by the Gaussian function. σ_j, the standard deviation of all non-zero values of the j feature is in the dataset for the second state a constant value of $-\lambda$ is considered irrespective of the value of the non-zero property value. For the latter, the attribute has no similarities (Taherizadeh and Moghadam, 2009).

The SMTP model to measure the similarity between two documents in textual processing is expressed as follows.

The definition of similarity is calculated using the following function:

$$F(d_1, d_2) = \frac{\sum_{j=1}^{M} N_*(d_{1j}, d_{2j})}{\sum_{j=1}^{M} N_{\cup}(d_{1j}, d_{2j})}$$

We define the numerator of the above function below:

$$N_{\cup}(d_{1j}, d_{2j}) = \begin{cases} 0.5\left(1 + \mathbf{EXP}\left\{-\left(\frac{d_{1j} - d_{2j}}{\sigma_j}\right)^2\right\}\right), & \text{For two text documents ofat least one similarity} \\ \mathbf{0}, & \text{Passwords are not found inany of the documents} \\ -\lambda & \text{Otherwise} \end{cases}$$

$$N_{\cup}(d_{1j}, d_{2j}) = \begin{cases} 0, & \text{If the keyword is not found in d1 and d2} \\ 1, & \text{Otherwise} \end{cases}$$

Then, using the following equation, the similarity between the two documents is calculated:

$$S_{SMTP}(d_1, d_2) = \frac{F(d_1, d_2) + \lambda}{1 + \lambda}$$

Euclidean distance

The second method is used to calculate the similarity between two Euclidean distances. The Euclidean distance between two sources of text can be explained as follows:

$$E_{EUC}(d_1, d_2) = \left[(d_1 - d_2) \cdot (d_1 - d_2)\right]^{1/2}$$

The two parameters d_2 and d_1 represent two text documents that are compared in the above relationship. In this study, these two methods have been used (Taherizadeh and Moghadam, 2009).

2.2.2 Davis boulder indicator

This criterion uses the similarity between two clusters (R_{ij}), which is defined based on the dispersion of a cluster (s_i) and the lack of similarity between two clusters (d_{ij}). Usually, the similarity between two clusters is defined as follows:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

d_{ij} and s_i are calculated with the following equations.

$$d_{ij} = d\left(v_i, v_j\right)$$

$$s_i = \frac{1}{\|c_i\|} \sum_{x \in c_i} d\left(x, v_i\right)$$

Given the stated content and the definition of the similarity between two clusters, the Davis Boulder index is defined as follows:

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i$$

$$R_i = \max\left(R_{ij}\right), i = 1, ..., n$$

This index actually calculates the average of the similarity between each cluster and the most similar cluster to it. It can be seen when the index is lower, better clusters are produced (Lin et al., 2014).

2.3 Multi-layer perceptron neural network

The perceptron algorithm was invented by Frank Rosenblatt in 1957 at the Aronothekan Kernel Laboratory. One of the most basic neural models available is the multi-layer perceptron model, or MLP, which simulates the transfusion function of the human brain. Each human brain neuron processes after receiving the input (from a neuronal or another non-neuronal) and transmits the result to another cell (neuronal or non-neuronal). This behaviour continues to yield a definite result, which will likely lead to decision making, processing, thinking or moving body parts. Usually a multi-layered perceptron neural network is composed of a set of sensory units. An input layer consists of one or more hidden layers of computational nodes and an output layer of computational nodes. The input signal propagates through the network in a forward direction layer by layer. Multi-layer perceptron is of forerunner artificial neural networks in such a way that each neuron output in one layer provides the input of other neurons. The multi-layer perceptron network uses a training method with monitoring. The purpose of training the network is to minimise the generated error, which is done based on network weighing.

Learning in perceptron

The main features of perceptron are the ability to learn or train. This learning is supervised in perceptron. In the sense that we need to have a number of inputs along with the correct output so that perceptron can mimic it. The learning process in perceptron is as follows:

- provide an output
- compare the output to the output that should be
- after repeating these steps sufficiently, the perceptron is converged.

2.4 PSO algorithm

PSO is an optimisation technique that acts based on a population of early responses. In this method, the system works with a population of a number of initial responses, and by moving these responses during consecutive repetitions it tries to find the optimal response (Das et al., 2008)). Among other evolutionary algorithms, PSO is more popular with ease of implementation and setting of several parameters. This algorithm has recently been used in the training of neural networks (Karwowski et al., 2013). An algorithm for optimising particle swarm in many areas, such as finding optimal solutions for functions, training nerve networks and so forth demonstrates its proper and acceptable performance.

3 Suggested method

In this section, the research methodology is first described. The data source and the word-centric search method are introduced respectively.

3.1 Research methodology

As stated, clustering has a steady role in web mining. In fact, by using the clustering technique, it is possible to categorise information and texts on the web for evaluation and production of knowledge. When monitoring the information on the web, the semantic meaning of words is not important. Sometimes and the only word presence in the text is sufficient to extract information. In other words, words have a singular meaning, and only the presence and use of them in the text by the user can transmit different information. In this situation, only word itself is considered not its meaning. A method for preparing web data with word-centric search is introduced. In 2017, Rezaeian et al. focused on the method of clustering K-means using the keyword in the research. In the present study which is inspired by Rezaeian et al. (2017) researches, information is clustered based on keywords in the web environment. So far, metrics for measuring similarity have been introduced to measure web clustering. In 2014, Lin et al. introduced SMTP similarity measurements. In this criterion, the similarity between the two documents is calculated based on the existence or non-existence of the feature. In this research, using selected keywords has been considered as a feature. The proposed system takes place in four steps:

- *Step one:* preparing data for web mining – word-centric search method

  Firstly the information block is identified. Then lines containing the text are separated, and after that a matrix called logic is created by specifying the keywords. This matrix represents the number of repetitions of each keyword in each block of information. At the end, the similarity of pairwise blocks of information and formation of the matrix degree are calculated by using the SMTP model which calculates the similarity between two documents based on the existence or non-existence of a keyword.

- *Second stage:* clustering using K-means algorithm at this point, information will be clustered by Euclidean distance and SMTP model. Now by using the matrix of the number of each keyword repetitions in each information block and the matrix of measuring the similarity of pairwise blocks of information, the K-means algorithm is used to cluster the information. The resulting output displays the cluster number of each block of information as a matrix. Then, these two criteria are compared by using the Davis Bouldin index which actually calculates the average of the similarity between each cluster with the most similar cluster.

- *Stage three:* implementing and performing the MLP neural network

  At this step of the research, the structure of the neural network will be determined. Then, the matrix considers the number of repetitions of each keyword in each block of information as the input of the neural network and the cluster number matrix of each block of information as the expected output of the neural network. After running the neural network, the output charts are analysed.

- *Step four:* neural network hybrid and PSO algorithm (PSO – MLP)

  In the final step which is the training stage, the researcher uses the PSO evolutionary algorithm to weigh the layers and examines the results to see whether a recovery in neural network performance is created.

3.2 Source of data

Every day the Mem Tracker Foundation checks more than 900,000 topics cited and searched by users on websites from over 1,000,000 websites, weblogs tabs in order to analyse information on how the news becomes viral and the length of time the posts related to that news were sent, forwarded and circulated on social networks such as Twitter. This is a very reliable source for analysts and information scientists and web mining.

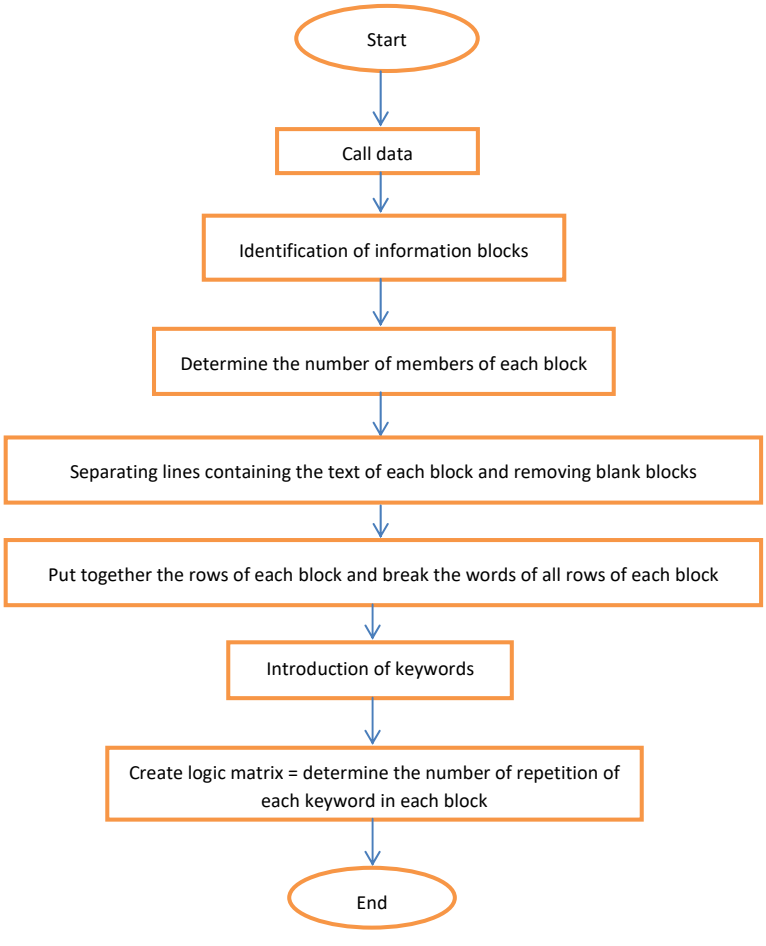
3.2.1 Data structure

Theoretically, data presented by four letters – P, Q, T, and L – at the beginning of each row expresses the following concepts.

Table 1 Data structure

P	Represents the site where the information has been published and is displayed standard (http//www.SOMEWHERE.com).
	P http://populargusts.blogspot.com/2009/04/walls-and-gates-of-seoul.html.
Q	A collection of words and sentences published in the source referred to previous row by users or newsgroups. Rows that begin with 'Q' can range between zero and several lines of content.
	Q the gate of bright amiability
T	The initial release date of the news or subject is based on the standard ****/**/** format.
	T 2009-04-01 00:00:06
L	Verifies other websites that published the news again by referring to the original source.
	L http://populargusts.blogspot.com/2007/06/century-old-books-about-korea.html
Length	Indicates the length of the information count on each row.

Figure 1 Flowchart of a word-centred search method (see online version for colours)



3.3 Word-centric search method

Step 1 First of all, the researcher uses step-by-step dedicated coding to identify the blocks of information. To achieve this, based on the startup index P, the number of blocks that started with this keyword is counted. The result of this count is the number of information blocks that include the items in the raw information form displayed. Therefore, each block of information extends from the keyword P to P. In general, the information blocks are the same as the site where the information has been published.

Step 2 After identifying the data blocks and counting them, it is necessary to proceed with the separation of the lines containing the text that will be used as the primary clustering data. Using the scan of rows in each block, the information begins with the letter Q. This text information is stored in dedicated cells in the exclusive number of each block, which can then be easily accessed.

Step 3 Definition of specific passwords that are used to compare and assignment of similarities. These passwords are defined by n and are stored in a 1 * n matrix.

Step 4 In this step, the number of each $n \in N$ is counted in each of the textual cells of the information blocks, the output of which is a matrix of 1 * n, the number of repetitions of each word in textual cell is displayed.

Step 5 Continue after step 4, for all the text cell blocks of the information, the researcher arrives at a matrix of a * b size, equal to the number of columns with the number of predefined passwords, and the number of rows of this matrix to the total number of text rows is the total data. This matrix is stored in the coding process called logic. The proposed method flowchart is shown in Figure 1.

4 Analysis of research findings

Information analysis is done to understand the facts and concepts. As discussed extensively earlier in the previous section, this paper attempts to provide a web-based method for clustering. Then the clustering of the data is given. Finally, the researcher focuses on using the neural network hybrid and the evolutionary algorithm.

4.1 Preparing data for web mining

4.1.1 Number of information blocks

According to the detailed format in the second part of this paper, the first step is to identify the method of separating information which is broken down according to the duplicate structure with unique passwords by using the coding in the MATLAB software environment, the information blocks, and for used raw information. These numbers are as follows. Table 2 is an overview of the number of rows per block.

Table 2 Number of rows per block

Information block number	*Number of rows*
1	6
2	7
3	11
4	9
5	5

4.1.2 Keyword introduction

At this part, the separation information blocks are being used according to the need or volume of the continuous of the words to select the keywords.

In this research, it is assumed that the selected keywords are defined according to the necessity of the organisation's need as follows:

$$Keyword = \{money, credit, bank, I, while\}$$

Using the proprietary function, the number of words in each of the blocks of information is identified. To generalise the coding process, the number of keywords is set according to the user's needs, so that the user can change the keywords and their number if needed.

The result of the above process is the production of a matrix with dimensions of 50×5 that contains the number of replicas of each keyword in each block of information stored with the logic symbol. Table 3 is a view of the logic matrix.

Table 3 The number of repetition of each keyword in each block of information

While	*I*	*Bank*	*Credit*	*Money*
0	1	0	0	1
0	0	0	0	1
0	4	0	0	0
0	4	0	0	0
0	0	0	0	1

4.2 Clustering using the definition of similarity

Based on the predefined steps and the results obtained from the word-centric search method stored in the logic matrix, the researcher clusters the data by using the Lloyd's algorithm and to introduce the distance in another word similarity the researcher uses Euclidean distance and SMTP model. The number of clusters is 5 which are justified due to the number of key words. The two criteria are compared by using Davis Boulder's index. Results are shown in Table 4. The DB value is lower which means better clusters are produced, thus the Euclidean distance criterion is better than the SMTP model.

Table 4 DB index results

Euclidean distance	*SMTP model*	*Method*
0.7501	2.4979	DB index

An illustration of the idx matrix, the cluster number of each block using the two criteria is given in Table 5.

Table 5 The cluster number of each block

Information block number	*Cluster number (Euclidean distance)*	*Cluster number (SMTP model)*
1	4	5
2	5	5
3	1	5
4	3	5
5	5	5

For better access, in the following, steps of the logic and idx matrix are stored in an independent matrix presented with the CUM symbol:

$$CUM = [logic, idx]$$

At this stage, using the sum of the referred propositions and the functions used, the clustering of the information has been completed and can be used in the neural network to continue work with the output of clustering using two criteria.

4.3 Define the structure of the neural network

Artificial neural networks are the proper tools for matching, learning and categorising information. In this paper, the neural network is studied in two independent and hybrid states using the particle swarm algorithm. At this stage, neural network and hybrid neural networks are performed based on two criteria.

- Neural network input data

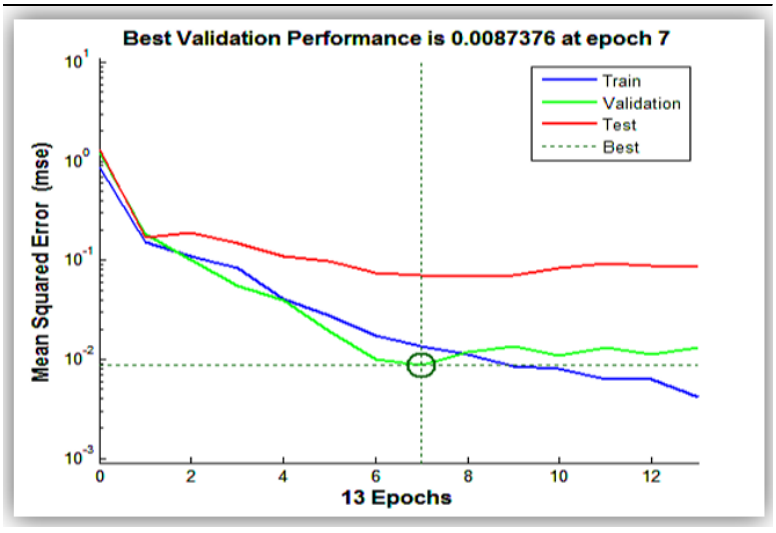
  In the previous section, the CUM matrix containing the primary transformed information was used and the cluster number accrued in this section.

  The neural network input contains all the rows and columns of the CUM matrix except the last column. Output demanded from the neural network, the last column of the above matrix, is also introduced. The two criteria difference is also in the demanded output from the neural network, because each of the two criteria has yielded different clustering results.

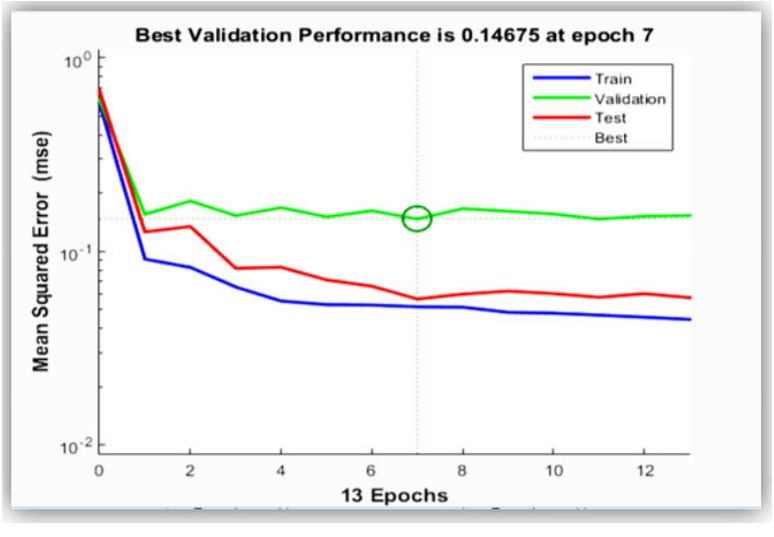
- Neural network structure

  Most researchers consider forerunner multi-layered neural networks, especially multi-layered perceptron networks, as one of the most successful clustering and forecasting networks. Therefore, in this thesis, a feed-forward neural network is used which is a form of multi-layer perceptron neural network, and the researcher constructs a neural network with ten secondary layers by defining NetPlain.

Figure 2 Function in specifying the model, (a) Euclidean distance (b) SMTP model (see online version for colours)



(a)



(b)

4.3.1 Neural network output with independent training

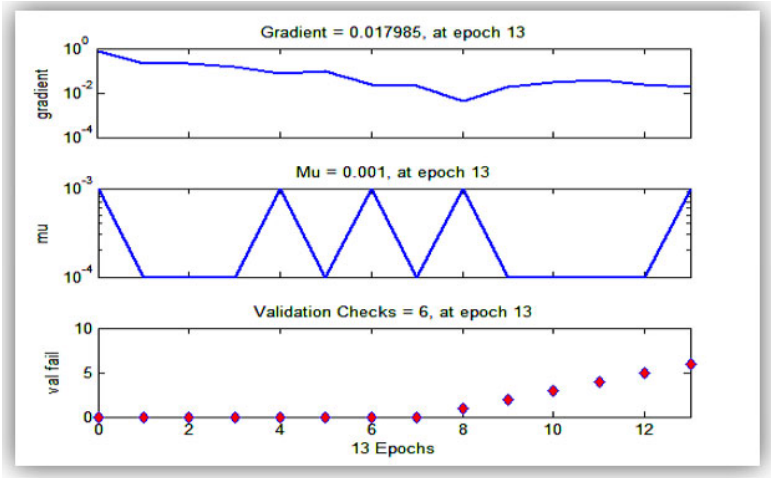
After calculating the weights of neurons and by using clustering output based on two criteria (Euclidean distance and SMTP model), expected input and output of neural network process was provided in Table 6. Regarding the comparison between the neural network according to the results of the two approaches, the root of the least squares of error is the training values based on the Euclidean distance of 0.1419 and based on the

SMTP model 0.2613. At this step, the neural network trained yield by the Euclidean distance results show the better results.

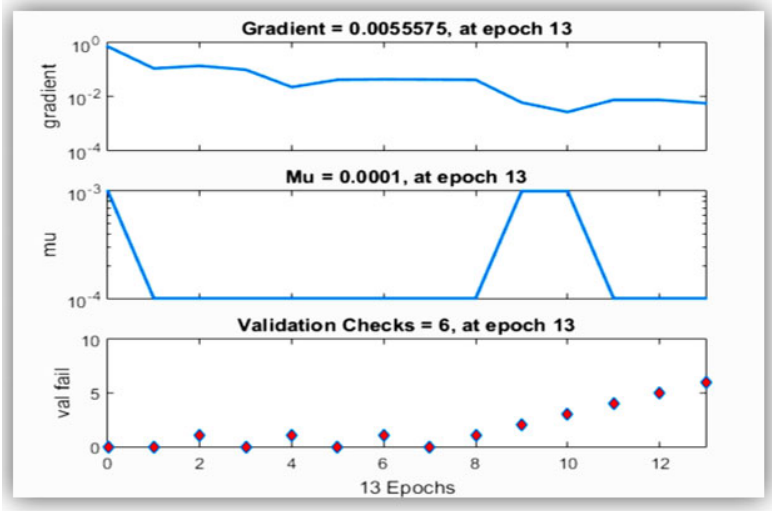
- The algorithm used to train the neural network: Levenberg-Marqauardt.
- Method of measuring the quality of modelling and fitting: average of the error sentences mean square error (MSE).
- The number of repetition using Euclidean distance criterion: 13 epoch.
- The number of repetitions using the SMTP model: 13 epoch.

The graphs of the output of the neural network are presented in Figures 2, 3 and 4.

Figure 3 Neural network training statistics, (a) Euclidean distance (b) SMTP model (see online version for colours)

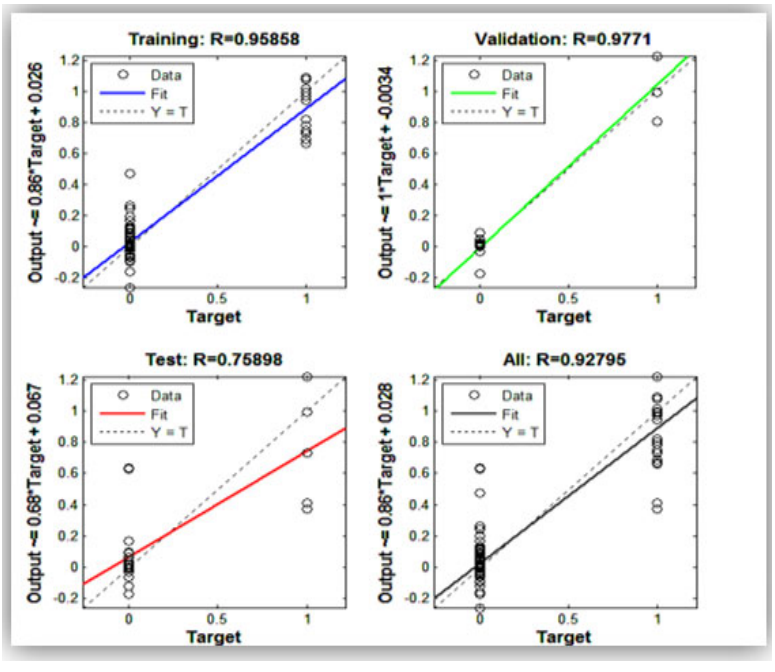


(a)

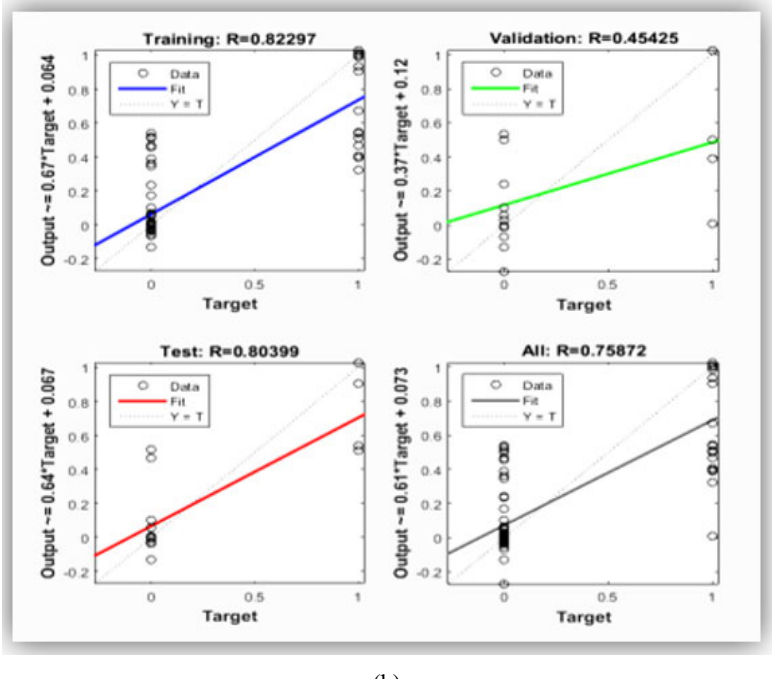


(b)

Figure 4 Regression equations, (a) Euclidean distance (b) SMTP model (see online version for colours)



(a)



(b)

Table 6 Average of squared error neural network

Parameter	*Explanation*	*Euclidean distance*	*SMTP model*
Average of squared error values of training	MSE_Train	0.0222	0.0683
Average of squared error test values	MSE_Test	0.0990	0.1020
Root of least squares error training values	RMSE	0.1491	0.2613
Percentage of error	Error percent	7.4550	13.065

Table 7 PSO algorithm settings

The number of repetitions	Iteration	20
Primary population volume	SwarmSize	100
Parameter C1	C1	2
Parameter C2	C2	4 – C1
Parameter C3	GlobalBest	inf

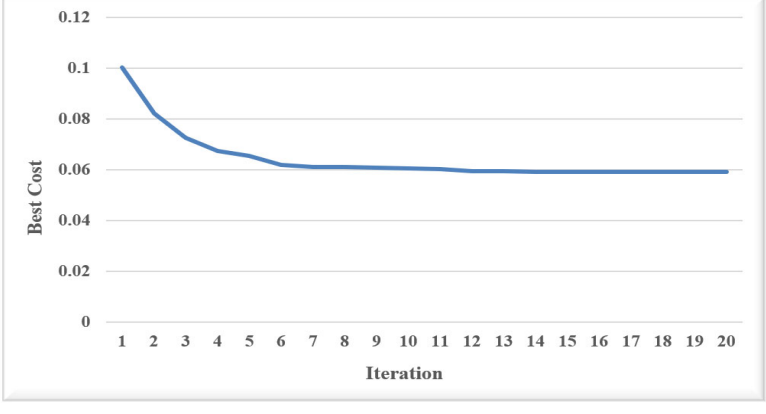
4.3.2 Hybrid neural network

In line with the research objectives at this step, the researcher addresses how the neural network responds to the application of the particle swarm algorithm in the training and initial weighing step and its effect on the final performance of the neural network. The settings for the PSO algorithm are given in Table 7.

Hybrid neural network output

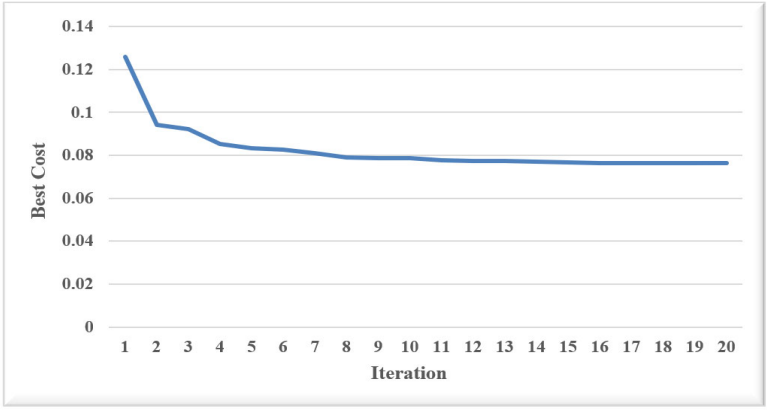
It should be noted that every time the program is executed, the numbers are different due to the randomness of the initial values of the weights, and running the program multiple times causes overtraining. The output of the hybrid neural network is shown in each illustration in Figure 5.

Figure 5 Best cost values per replication, (a) Euclidean distance (b) SMTP model (see online version for colours)



(a)

Figure 5 Best cost values per replication, (a) Euclidean distance (b) SMTP model (continued) (see online version for colours)



(b)

5 Conclusions and suggestions

The present research provides a method to prepare web data for clustering, under the title word search. Since the neural network is used for prediction and clustering, for the training of the neural network, data should be determined as the expected output, so first the web data is clustered using the K-means based on two Euclidean distance and SMTP model criteria. According to the DB, the Euclidean distance criterion has better performance in clustering. Then the results of clustering are considered as the expected output of the neural network. Neural network and MLP-PSO hybrid function were investigated in two cases. Based on the Lloyd's algorithm used in this study, several quantitative criteria can be applied to calculate distances, which is unlikely to get the exact same results or results that will confirm them by changing the approach from the Euclidean distance to the city block distance which did not happen practically. The clustering results obtained from two Euclidean and SMTP models were different. Also, by changing the data and volume and span of it, it is expected that there are completely independent responses, because, by changing the data, all the effective parameters in the cluster will encounter fundamental changes.

A change in the keywords can face the whole output with the most fundamental changes. This sensitivity to the clustering of the text data and in another word, sensitivity towards the keywords and the type of data can be further developed and studied by other researchers. The word-centric search method can be used in various fields. For example, by searching for various news sites, it can be found out that which words have repeated on the various sites. Therefore, it can be concluded that each day's news is mostly about what topic. Also, to measure the incidence of a business, you can use this method to search for repetitions of your business name and other competitors on web pages and user posts.

Regarding the comparison between the neural network according to the results of the two approaches, the root of the least squares of error is the training values based on the

Euclidean distance of 0.1419 and based on the SMTP model 0.2613. At this step, the neural network trained yield by the Euclidean distance results show the better results.

With regard to the comparison between the neural network and hybrid neural network of the perceptron and the PSO algorithm, the network was run in accordance with the requirements of the lack of overtraining of the network in both modes. The results are shown in Table 8.

Table 8 Comparison of MLP and MLP-PSO results of Euclidean distance and SMTP model

Parameter	*Explanation*	*Euclidean distance*		*SMTP model*	
		MLP	*MLP-PSO*	*MLP*	*MLP-PSO*
Average of least squares error training values	MSE_Train	0.0222	0.0763	0.0683	0.0590
Average of least squares error test values	MSE_Test	0.0990	0.1089	0.1020	0.0917
Root of least square error	RMSE	0.1491	0.2763	0.2613	0.2430
Percentage of error	Error percent	7.4550	13.8137	13.065	12.1084

Based on average criterion of the square error of the training values, the results are as follows:

- based on the Euclidean distance criterion, neural network training presents better results than neural network hybrids
- based on the SMTP model, neural network hybrid training presents better results compared to the neural network.

After implementing the program for several times, the results of the unweight average are given in Table 9. The results show that the MLP-PSO hybrid performs better based on the SMTP model.

Table 9 Results of the yield of un-weighted average MLP-PSO hybrid

Parameter	*Explanation*	*Euclidean distance*	*SMTP model*
Average of least squares error training values	MSE_Train	0.09416	0.06664
Average of least squares error test values	MSE_Test	0.12172	0.0971
Root of least square error	RMSE	0.3062	0.2579
Percentage of error	Error percent	15.2986	12.8894

Some suggestions for further studies are:

- using the proposed method for classifying web data by using neural network hybrid with different evolutionary algorithms
- due to the fact that the web data of this study is not related to a particular topic, this method can be used in a particular area such as news sites, sports, brand, etc., and its performance in clustering of the data is analysed.

References

Borzemski, L. (2006) 'The use of data mining to predict web performance', *Cybernetics and Systems: An International Journal*, Vol. 37, No. 6, pp.587–608.

Chou, P.H., Li, P.H., Chen, K.K. and Wu, M.J. (2010) 'Integrating web mining and neural network for personalized e-commerce automatic service', *Expert Systems with Applications*, Vol. 37, No. 4, pp.2898–2910.

Cooley, R., Mobasher, B. and Srivastava, J. (1997) 'Web mining: Information and pattern discovery on the World Wide web', in *Proceedings, Ninth IEEE International Conference on Tools with Artificial Intelligence*, IEEE, November, pp.558–567.

Das, S., Abraham, A. and Konar, A. (2008) 'Automatic clustering using an improved differential evolution algorithm', *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, Vol. 38, No. 1, pp.218–237.

Etzioni, O. (1996) 'The World-Wide Web: quagmire or gold mine?', *Communications of the ACM*, Vol. 39, No. 11, pp.65–68.

Fürnkranz, J. (2005) 'Web mining', in *Data Mining and Knowledge Discovery Handbook*, pp.899–920, Springer, USA.

Gedov, V., Stolz, C., Neuneier, R., Skubacz, M. and Seipel, D. (2004) 'Matching web site structure and content', in *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, ACM, May, pp.286–287.

Jain, A.K. (2010) 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters*, Vol. 31, No. 8, pp.651–666.

Jokar, N., Honarvar, A.R., Aghamirzadeh, S. and Esfandiari, K. (2016) 'Web mining and web usage mining techniques', *Bulletin de la Société des Sciences de Liège*, Vol. 85, No. 1, pp.321–328.

Karwowski, J., Okulewicz, M. and Legierski, J. (2013) 'Application of particle swarm optimization algorithm to neural network training process in the localization of the mobile terminal', *Proceedings of International Conference on Engineering Applications in Neural Networks*, pp.122–131.

Kazienko, P. and Kiewra, M. (2003) 'Link recommendation method based on web content and usage mining', in *Intelligent Information Processing and Web Mining*, pp.529–533, Springer, Berlin, Heidelberg.

Kosala, R. and Blockeel, H. (2000) 'Web mining research: a survey', *ACM Sigkdd Explorations Newsletter*, Vol. 2, No. 1, pp.1–15.

Lin, Y.S., Jiang, J.Y. and Lee, S.J. (2014) 'A similarity measure for text classification and clustering', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 7, pp.1575–1590.

Momeni, E., Armaghani, D.J., Hajihassani, M. and Amin, M.F.M. (2015) 'Prediction of uniaxial compressive strength of rock samples using hybrid particle swarm optimization-based artificial neural networks', *Measurement*, Vol. 60, pp.50–63.

Orhan, U., Hekim, M. and Ozer, M. (2011) 'EEG signals classification using the K-means clustering and a multilayer perceptron neural network model', *Expert Systems with Applications*, Vol. 38, No. 10, pp.13475–13481.

Rezaeian, M., Montazeri, H. and Loonen, R.C.G.M. (2017) 'Science foresight using life-cycle analysis, text mining and clustering: a case study on natural ventilation', *Technological Forecasting and Social Change*, Vol. 118, pp.270–280.

Srivastava, J. and Mobasher, B. (1997a) 'Panel discussion on 'web mining: hype or reality?'', in *The 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 1997)*, Newport Beach, CA.

Srivastava, J. and Mobasher, B. (1997b) 'Web mining: hype or reality', in *9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, Vol. 275, p.282.

Taherizadeh, S. and Moghadam, N. (2009) 'Integrating web content mining into web usage mining for finding patterns and predicting users' behaviors', *International Journal of Information Science & Management*, Vol. 7, No. 1, pp.51–66.

Wang, X., Abraham, A. and Smith, K.A. (2005) 'Intelligent web traffic mining and analysis', *Journal of Network and Computer Applications*, Vol. 28, No. 2, pp.147–165.

Za'in, C., Pratama, M., Lughofer, E. and Anavatti, S.G. (2017) 'Evolving type-2 web news mining', *Applied Soft Computing*, Vol. 54, pp.200–220.

Zhang, J.R., Zhang, J., Lok, T.M. and Lyu, M.R. (2007) 'A hybrid particle swarm optimization – back-propagation algorithm for feedforward neural network training', *Applied Mathematics and Computation*, Vol. 185, No. 2, pp.1026–1037.

Websites

https://www.acm.org/.

https://www.siam.org/.

Notes

1 Feedforward neural network.

2 Adaptive particle swarm optimisation algorithm.

3 Knowledge discovery database.