Journal of Web Engineering, Vol. 13, No.3&4 (2014) 333--360
© Rinton Press

WEBPAGE CLUSTERING – TAKING THE ZERO STEP: A CASE STUDY OF AN IRANIAN WEBSITE

ABBAS KERAMATI RUHOLLA JAFARI-MARANDI

School of Industrial and Systems Engineering, College of Engineering, University of Tehran, Iran
Keramati@ut.ac.ir ruholla.jafari@ut.ac.ir

Received December 4, 2013
Revised April 4, 2014

The expansion of websites and their too many pages not only have pushed their visitors to frustration but also have made the websites ever more difficult to be managed and controlled by their owners. In the past few years data mining (clustering) has been of great help so as to assist website's owner to address the complexities related to owners' extracting their visitor's preferences and their coming to know their websites properly. In this line of literature, this paper contains several parts and features. First, with regard to the fact that SOM has been the popular algorithm in dealing with page clustering, a comparison between SOM and K-means (another popular clustering algorithm) were performed to show the superiority of SOM in dealing with the task of webpage clustering. Second, due to the clustering tasks' complication not being able to be tested (unlike Classification), this study aims at proposing a mind-set by which one before taking any other actions has to go through some steps in order to choose the best set of data. Thirdly, looking at the literature, one can see the question about the suitability of types of data (content, structure and usage) and the task they are being used for has never been raised. Using an Iranian website's data, a field study and SOM algorithm, we presented that the popular belief about the type of data and the task they are appropriate for should be open to doubt. It was also depicted that different sets of data in two chosen tasks – webpage profiling and extracting visitors' preference - can influence the results tremendously. Last but not least, apart from observing the influence of different sets of data, both data mining tasks have been performed to the end and the results are presented in the paper. Additionally, using the second clustering task's results (the extraction of visitors' preferences) a novel recommendation system is presented. The recommendation system in question was installed in the website for more than a month and its influence on the whole website is observed and analysed.

Key words: Webpage clustering, Self-Organizing Map (SOM), K-Means, Recommendation System, Content Data, Structure Data, Usage Data
Communicated by: M. Gaedke & M. Bieber

1 Introduction

Nowadays websites' owner(s) have started to merit their visitors as potential customers whose visits can bring about lots of money. However, this is mostly the case for company or enterprise websites whose principal way of earning money is directly connected to their customer spending [1]. Even before these websites, there are business and websites in the World Wide Web (WWW) whose only reason of existence is because of their visitors. To put it another way, for world-wide-known websites such as Facebook or even Google the only reason behind their existence and working is their users (visitors). Being visited by a significant number of visitors, these websites grasp every opportunity to make more profits. Therefore, today the front line of the war on the internet can be summarized into three words which are “having more visitors”.

Over the years by expansion of WWW we have reached to a point that achieving this goal, having more visitors, does not seem so simple anymore. As to address these complexities and to survive in the bloody battle of attracting and holding on to their visitors, a significant amount of websites owners’ and researchers’ attentions have been drawn to data mining techniques. Adaption of different data mining techniques for different website’s task and problems has been the subjects of many researches in the past few decades. In general the term “Web Mining” is used for applying data mining in web and has three distinct areas [2]: web content mining, web structure mining and web usage mining. Web content mining is related to the web’s content such as text, image, audio, video, metadata, and hypertexts and the effort is to extract useful concepts and rules and summarize the content on the web. Whereas, web structure mining is related to underlying link structures of the web and its aim is to categorize webpages, measure similarities and reveal relationships between different Websites. Web usage mining is related to Web users’ interactions with the web and its aims are to extract patterns and trends in the users’ behaviors.

One of the very first step in applying data mining techniques to real world business or commercial problem is the recognition of business problem [3]. However, this fact necessitates the recognition of the business itself which is only possible, in the most cases, for an expert inside the business and not a newly-recruited Data miner. Now for a recognition of a website problem, coming to know a website is of the website owner’s or its data miner’s interest. Fortunately, if we are to take wise steps, data mining has many techniques to offer, not just to address different sites’ problem or task, but to gain information and knowledge about the sites and also to better our understanding of them, and consequently to classify and categorized these gained information to use for future data mining or other purposes. However, performing the tasks such as summarizing, profiling and visualization on a website are not small tasks and can be broken into three categories: website structure [4], website content [5] , and website visits [6].

With the growth of websites these task have ever become more complicated. It is not exaggerating to say that the website owners themselves do not know their own website. In another world, each website is a world by itself. Coming to know the magnitude and different angels of a website can be a challenging task due to their inexorable growing. Webpage clustering is an approach which can be used for approaching many cases and reaching answers for many other issues. As mentioned earlier, it can be employed to shed light on website for the owners to see their websites different angles. Webpage clustering, also, can be used to help the process of helping the visitors to find what they desire. There are many issues and concerns involved in what should be done for the task to be successful, such as the selection of data, the algorithm and the process should be taken. However, this paper follows two main streams. Its first effort is to show the importance of selecting appropriate sets of data according to the task in had - a webpage clustering task. Second, we desire to introduce a necessary zero step in the process of webpage clustering. Additionally, several other assumptions and claims were observed, tested and analysed, among others we can mention presenting 1) the superiority of SOM over K-Means algorithm in time of dealing with clustering task, 2) the elimination of a popular belief that every types of data is appropriate for a certain task, for example, usage data is assumed appropriate for the extraction of visitor preferences, 3) (only in this paper case) the fitness of the combination of structure and content data for webpage profiling and also the fitness of the combination of usage and content data for visitors' preference extraction, 4) the assumption that using hexagonal or quadrilateral do not have significant influence on SOM performance in dealing with webpage clustering, and 5) the positive influence of employing a clustering based recommendation system on our case website.

The rest of this paper is presented as follow. Section 2 is literature review which represents related studies and efforts, whereas section3 introduce the two techniques used in this study, namely Self-Organizing Map (SOM) and K-Means algorithms. Section 4 is the main part of this paper which starts from presenting assumptions and end with the discussion of the results. Finally, section 5 is for the conclusion, applicability and future trend discussions.

2 Literature Review

The literature of webpage clustering, or document clustering in the first place, goes back to the work of Kohonen et al. [7] whose aim was to propose a system that would be able to organize vast document collections according to textual similarities. Apart from the data mining technique applied to the problem, that is SOM, the only input data for the task has been the content (textual data) of the documents. However, in case of webpage clustering there are more data available beside the content. For one, there is the data of user click stream which can be used to cluster webpages and also the data of relation and structure of these webpages can be another source of data for the task. Looking at the related literature, one can see that the source of data for the task is also parted into three categories: clustering by the data of (1) webpage content, (2) webpages structures, and (3) click streams and visitors' behavior (usage).

Studies which addressed webpage clustering using webpage content or web page structures data are easily outnumbered by those that used visitor behavior data. This fact may just be due to the attractiveness of visitor data. However, as it was mentioned earlier, Kohonen et al. [7] used textual similarities to cluster a pile of documents. They used statistical representations of their vocabularies as the feature vectors for the documents and this was, as they cited, the reason why Self Organizing Map (SOM) algorithm was employed. Incidentally, one of their claimed contributions was to adapt SOM and better its performance for the problem. In a similar study Huang et al. [8] employed the unsupervised algorithm in order to cluster Chinese patent documents. On the contrary to former study, the latter took advantage of both the structure and content data of document in the process of clustering. Their intended difference between structure and content data was that the former is the data of the relation between documents whereas the latter is the independent data of each document.

The application of user behavior data, or web usage mining, with the view of coming to know website employing a clustering algorithm has been under the spot light over the few past years starting by Su et al. [9]. Although the study main aim is to propose an effective approach for adapting web interfaces to improve visitor's experience, the principal means to do so has been clearly webpage clustering using web logs. By introducing index pages whose aim is to minimize overall user browsing, an automatic method for web interface adaptation was presented. The effort might be recognized as the pioneer to employ clustering webpages in order to improve website adaptation, however, Recursive Density Based Clustering (RDBC) algorithm was applied. In an interesting study by Ypma et al. [10] one more step was taken, and the knowledge of clustered visitors was used to categorize webpages. However, unlike the previous study, they trained and exploit of a mixture of hidden Markov model which is, of course, not a common technique in the literature.

Clustering web pages using the visitors' behavior with the means of Self-Organizing Map (SOM) has been prevalent in the related field, see Table 1. Smith and Ng [5] are among these researchers in whose work SOM was used to mine web log data in order to provide a visual tool with the view to assisting user navigation. A LOGSOM was adapted, from traditional SOM, to organize web pages into a two-dimensional map. Qi and Li [11], and Kim [12] are the authors of another two studies with a same theme. While the former's main objective is to help understanding visitor behavior by the means of clustering algorithm, SOM and K-Means, the latter's aim is to understand the visitor navigation by grouping them and consequently visualizing and clustering webpages for further purposes. In addition, Park et al. [2] applied adapted SOM, however, to cluster website's visitor to web logs data. Moreover, this study main focus, apart from clustering visitors, has been the improvement of the algorithm.

Lin and Tseng's [4] rather different perspective to other researcher was to shed light on the website organization itself instead of on sub-systems such as recommendation. That is to say, this study objective is to use web mining to restructure website organization in order to improve it for the better experience of visitors. Furthermore, Tsekouras et al. [13] proposed a categorical data fuzzy clustering algorithm to classify web documents. The interestingness of this study is ascribed to its fuzzy approach. However, the study should be categorized into webpage clustering which uses webpages content data.

Table 1 Literature Review Summary

		Type	Method	Data	Task
1	(Park, et al., 2008) [2]	Web usage mining	statistical analysis, clustering, classification, association rules, sequential pattern discovery, and dependency modelling	Web server logs - User navigation patterns	Clustering web users
2	(Smith & Ng, 2003) [5]	Web usage mining	SOM	Web server logs - User navigation patterns	Crusting webpages
3	(Kim, 2007) [12]	Web usage mining	Adapted SOM – visualization	User navigation patterns	Clustering webpages
4	(Huang, et al., 2008) [8]	Web content and structure mining	SOM	Chines patent documents	Clustering books
5	(Qi & Li) [11]	Web usage mining	SOM	Web logs	Clustering webpages
6	(T. Kohonen, et al., 2000) [7]	Web content mining (textual similarity)	Scaled up SOM	Textual	Clustering document
7	(Su, et al., 2002) [9]	Web usage mining	Recursive density based clustering algorithm	Web logs	Clustering webpages
8	(Ypma, et al., 2002) [10]	Web usage mining	Hidden Markov model Clustering	Web logs	Cluster webpages
9	(Tsekouras, et al., 2007) [13]	Web content mining	Categorical data clustering (CDC)	Downloading pages and extracting words	Cluster web documents
10	(Lin & Tseng, 2010) [4]	Web structure mining	Ant colony	--	Website reorganization

Looking briefly at the literature (Table 1) two common issues for all of the works are outstanding. First, SOM was the most prevalent technique applied to address the task. Secondly, it is noticeable that most of the researchers have been keen on clustering the pages or the documents based on the data they had at their disposal and they failed to understand the importance of the different types of data, and to make use of the data in hand wisely. As it was presented there are three known types of data: content, structure and usage data. Each and every one of them has completely different nature which, in turns, spells their different probable usage. To the best of our knowledge there has never been a study in order to investigate through the probable different influence of different types of data. However, this study's main aim has been to propose a mind-set by which one has to select the best set of data before performing any clustering task. Additionally, this paper studies a real website, distinguishes between its different kind of data, proposes a zero-step methodology to find the best set of data, applies the methodology for two different tasks (webpage profiling and extracting visitors' preferences), and finally goes on to perform the tasks and install them on the website based on the methodology's results to oversee the influences.

3 Techniques

3.1. Self-Organizing Map (SOM)

Self-Organizing Map (SOM) is a technique which falls into the category of Artificial Neural Networks. Known also as Kohonen Networks, named after its inventor Dr. Tuevo Kohonen [14], it is mostly used for undirected or unsupervised tasks. That is to say, on the contrary to usual Neural Networks which are used for directed task such as predication or classification, the SOM type of Neural Network is appropriate for unsupervised task such as cluster detection. Although they bear some resemblance to one another and are based on the same underlying units as feed-forward, and even back propagation networks, SOMs are differentiable in two respects. Firstly, they possess a different topology and the back propagation method of learning is not applicable. Secondly, their method of training is completely distinguishable from that of feed-forward Networks. [15]

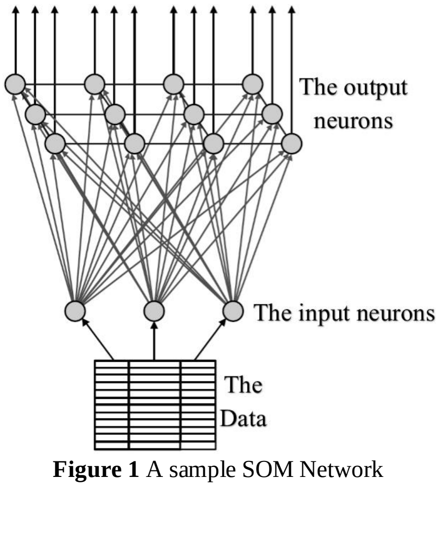


Figure 1 A sample SOM Network

A SOM Network adapted for a dataset with three attribute, accordingly three input neuron, and with 12 output neurons is presented in Figure 1. As one can see in the figure there must be a total number of input neurons as many as the dataset's attribute. However, the number of output neurons and their topology are variables by which SOMs can be tuned. The output layer consists of many units instead of just a handful. Each of the neurons in the output layer is connected to all of the neurons in the input layer. Conversely, not all the neurons are connected to each other in the output layer but only to their neighbors. Each neuron in output layer may have different numbers of neighbors with respect to the SOM's topology. In the figure each one has at most four neighbors due to its rectangular topology.

Tan et al. [16] Put SOM under the category of prototype-based clustering. In a prototype-base cluster any object is most similar to the prototype of the cluster than any other cluster's prototype. To make it more clear, in this point of view each output neuron is seen as the prototype of each cluster. The well-known K-Means is a prototype-based clustering and its prototypes are the centroids of the clusters. Similar to K-means, SOM process object one at a time and the centroid of the associated cluster is updated; contrary to K-means, a topographic neighboring is imposed upon the clusters in SOM which will bring about the neighbors of the associated cluster having to be also updated. To emphasize, the most compelling difference of SOM from other prototype-based clustering is that centroid (prototype) used in SOM have a predetermined topographic ordering relationships. In the course of training, SOM uses entering object to update the closest centroid and its neighbors. Adapted from [16], at a high level SOM clustering can be performed using the flowchart depicted in Figure 2.

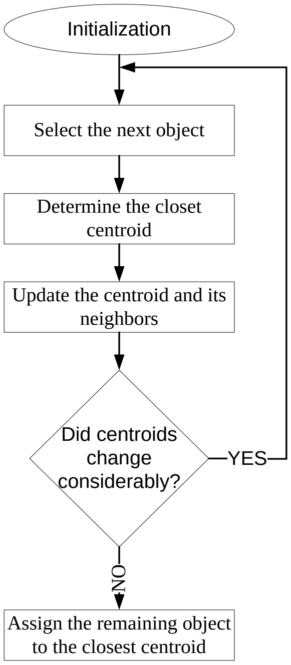

Initialization
Select the next object
Determine the closet centroid
Update the centroid and its neighbors
Did centroids change considerably?
YES
NO
Assign the remaining object to the closest centroid

Figure 2 SOM Flowchart (adapted from Tan et al. 2007) [16]

To explain each step of the flowchart (Figure 2), initialization is the act of specifying each neuron a centroid. They can be assigned with the randomly selected values from the range of attributes' values, or it is possible, similar to K-Means, to choose arbitrarily from the objects to be the centroids. The latter is more efficient and has less time-consuming nature [16]. Once the initialization is over, the time arrives to select next object so as to update the centroid. The task is pretty straightforward; selecting objects form training set which simply can be done randomly. Nonetheless, there are complications and issues about this step too. A dataset with small number of objects will spell the undergoing of each object more than once, whereas not all the objects of a large dataset are needed to be used. Similarly, the step of determining the closest centroid is not challenging except for choosing a distant metrics, which in the most cases is Euclidean.

However, the updating step is the most demanding and abstruse step to grasp. To update the closest centroid with the new-coming object the following equation can be employed.

$$m_i^{t+1} = m_i^t + h_i^t \times (p^t - m_i^t) \tag{1}$$

In this equation, m_t^i is the i^{th} centroid in the t^{th} step, p^t is the current new-coming object in the t^{th} step, and h_t^i is known as mediator coefficient and serves two purposes. Firstly, diminishing with time, it will gradually freeze the clustering procedure. Second, the coefficient imposes a neighborhood effect on the centroid nearest to the m_t^i. For the purpose of future reference, we know m_k^t as the centroid which is the nearest to the p^t in the t^{th} step. However, two different functions are often used in order to simulate the diminishing and imposing effect for h_t^i – Gaussian and step function. By contemplating the essence of h_t^i the imposing nature of the mediator on the neighbors will be apparent. The two functions are as follow.

$$h_i^t = \alpha(t)\times\exp(\frac{-dis(r_i - r_k)^2}{2\sigma^2 t}) \tag{2}$$

$$h_i^t = \begin{cases} \alpha(t) & dis(r_i - r_k) \le treshhold \\ 0 & otherwise \end{cases} \tag{3}$$

Both functions include *α(t)* and *dis(r_i-r_k)*. The former is a learning parameter whose value is between zero and one and diminishes with time to control the rate of convergence. Whereas, the latter is the distance between two centroid: r_k is the coordination of m_k^t and r_i is the coordination of centroid commensurate to h_i. The distance serves the purpose that the closest centroid to the m_k^t will undergo the most abrupt change, whereas the furthest will suffer the least change. Moreover, both functions will manipulate the width of the neighborhood by different approaches. Gaussian, Eq (2), uses the δ whose small value yields a narrow neighborhood and vice versa. Step function, Eq (3), has also contrived a threshold to meet the similar end.

As depicted in the flowchart, the next issue is when the loop should terminate. Ideally the loop should continue until there is a compelling evidence that a convergence has occurred: either no or little change in centroid vectors. And finally, when the convergence is met and all the data has been used in the course of training, the remaining data need to be assigned with the nearest centroid.

There are some criteria and variable which can affect the SOM demeanor dramatically. First, contriving a good start can make a whole difference in the performance of the technique. Second, the orders the data come to have the centroids of neurons updated can have a significant influence on the outcome. Last but not least, the distance measure, the mediator coefficient and the stoppage criteria may be manipulated to tune the technique. It is noteworthy that the most often used transformation for the SOM is the act of mapping the attributes between -1 and +1 due to the fact that all of the attributes must have the same influence [16]. We used the same transformation technique for all of our clustering practice with SOM.

3.2. K-Means

It is hard to defy the claim that one of the most in-used, popular and easy-to-apply clustering algorithm is k-means. The algorithm generates k points as initial centroids randomly. K is the number of clusters specified by the user. This last point, however, is recognized as the biggest disadvantage of the algorithm since it cannot find the optimum number of clusters on its own. Anyhow, after the initialization of the k centroids, each point will be assigned to the closest centroid and thus forming K clusters. Afterwards, the centroids will be updated based on the features of members' of clusters and afterwards some data points may move from one cluster to another. This, updating centroids and moving members, will go on until no change is possible or a convergence criteria is met. Like most of other prototyped clustering techniques, in this algorithm Euclidean distance is also used so as to find distance between data points and centroids. A flowchart for the algorithm is presented by Figure 3.

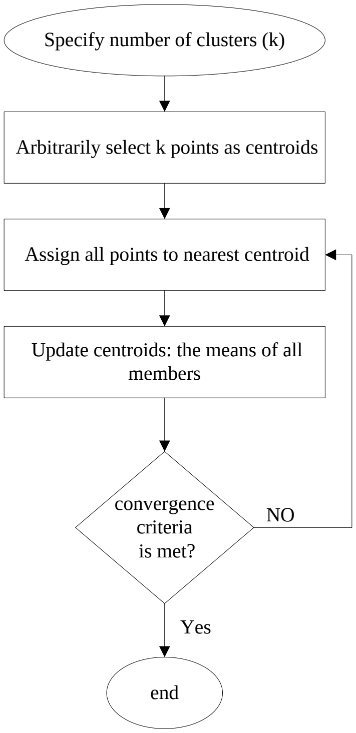

Specify number of clusters (k)
Arbitrarily select k points as centroids
Assign all points to nearest centroid
Update centroids: the means of all members
convergence criteria is met?
NO
Yes
end

Figure 3 K-Means Flowchart

As one can see the algorithm is understandable and easy to be grasped. However, the only issue remaining to be elaborated is the stoppage criteria. As pointed out by Tan et.al. [16], there can be different objectives for better convergence of clusters: 1) minimizing the sum of distances between points and assigned centroid, 2) minimizing the sum of squared distances between points and their assigned centroid and 3) maximizing the cosine similarity between points and their assigned centroid. These objectives are used with regard to the task that K-means is used. On average, if the complete convergence, i.e. when centroids do not change in iterations, is not possible, an objective with a criteria value will be used to stop the algorithm appropriately.

4 Research framework

The process and assumptions by which we shows the necessity of taking a prerequisite step for webpage clustering task is presented as follows. First, we run some experience to show the superiority of SOM in approaching a webpage clustering task. Although one could saw in the literature that nearly all of the researchers preferred SOM, we compared SOM with famous K-means to observe its superiority. Second, so as to validate our assumption that choosing different kind of data for approaching a webpage clustering is important we picked two different tasks: webpage profiling and webpage clustering based on visitors' preferences. Third, we took advantage of the data of a real website – Khatekhalagh. Forth, to effectively show the influence of data setting on the results we, aside from the data used for the two tasks, contrived two sets of ground information for the same website. These two sets of ground data were used to cluster the records twice. The results were compared and contrasted with the results from the two tasks using different sets of data: Content, Structure, and usage. Finally, we continued to perform the webpage profiling task using the set of data chosen from previous step to see the actual result. We also installed a recommendation system based on the results of our experience with the visitors' preference extraction and oversaw its influence on the success of the website on the whole.

4.1. The Website

Khatekhalagh[a] is a website that falls into the category of those with entertaining and amusing content. While it cannot be counted as a complete success, the site has been growing and improving in few past years. Given that the website lifetime has been less than two years in which time its owner hasn't had any financial support, having an Alexa rank under 100,000 can be considered as an achievement. Nevertheless, Khatekhalagh's owner in the time of constructing the website had in mind they might come to use their visitor click stream data and that lead them to keep each and every click of their visitors as a record. This advantages and their enthusiasm to blossom made it possible for, probably, the very first data mining study among Iranian websites. However as this study's aim and also the first step of data mining is business or website recognition, in this paper one of the efforts is to come to know the website better using the data mining itself.

As it was mentioned, the website main purpose is to represent entertaining content for its visitors. The structure of the website is very similar to others Iranian websites. However, it has some differences as well. To give an overall picture, it has four types of pages: (1) the entrance and main page which is the representation of latest posts' link, category pages' link and the most favorable posts' link; (2) category pages which are once again the representation of the latest posts and the most favorable posts' link in the particular category and also contains the tags with the most posts in that category connected to them; (3) Post pages which are the representation of the actual interesting content with some other posts' link suggestions: mainly the posts that has been tagged with the same words; and (4) the tag pages which apart from the posts which has been tagged with the word that is associated to the tag page represent the most favorable posts' link too. The small overview of these four page types is represented in Table 2. In the table the two last columns are respectively the number of visits from each type page and the number of advertisement visits from each type page. However, it is noteworthy that every page has two advertisement banners on the top-left corner, and also every post is related to at least one category page and one tag word.

Table 2 Khatekhalagh overall views in 2013, April 14

	Numbers of pages	Numbers of visit in the one month	Number of advertisement visit in on month
Main page	1	20841	84
Categories	30	16968	54
Posts	1090	251183	384
Tags	7812	167237	195
Total	8933	456229	717

[a] http://archive.khatekhalagh.com

As the effort of this study is to propose a model for better recognition of a website's content, in this case study our main purpose is to employ profiling, clustering and other data mining techniques to cluster the webpages of the website. The principal contents of the website are represented in its posts page. So the case study will be reduced to the clustering of the post pages with the help of available data. Fortunately, in case of this website we had access to the three types of data: website content data, website structure data, and visitor behavior data (website usage data).

4.2. Database and data selection

One of the most important and demanding part of any data mining task is data selection. In order to perform an efficient data selection one not only should be familiarized with the concept of the task and the business, here a website, but they must be ready to face limitation and privacy issues too. Not all of the data a data miner desire to have are being, or have been, produced; therefore, they must try to exploit the best of what is available. To make it more clear, Figure 4 is all of the tables we managed to gain access out of the website's database. There are, for sure, more data in database; however, it is so common for a data miner not to have access to all of data in databases. Moreover, there are the cases which a table or a simple query is provided for a data minder, especially for research purposes, which will force us to make do and bend with what we are provided.

Fortunately, the tables we were provided contained the three types of data (Figure 4): Content, Structure and usage. Table posts can provide us with valuable content data about posts, whereas the two tables – Tags and Categories – are suitable to extract some structural data. Last but not least, the table PageVisist is actually the practical means of collecting usage data. There might be different ways of recoding usage data. However, for this particular one keeping a record of each and every page visit in the table is the answer to the need. That is to say, when anyone visits one of the pages of this website, the website will record this as a record, which contains various types of information, and this will make the database prolific enough to extract usage data as well.

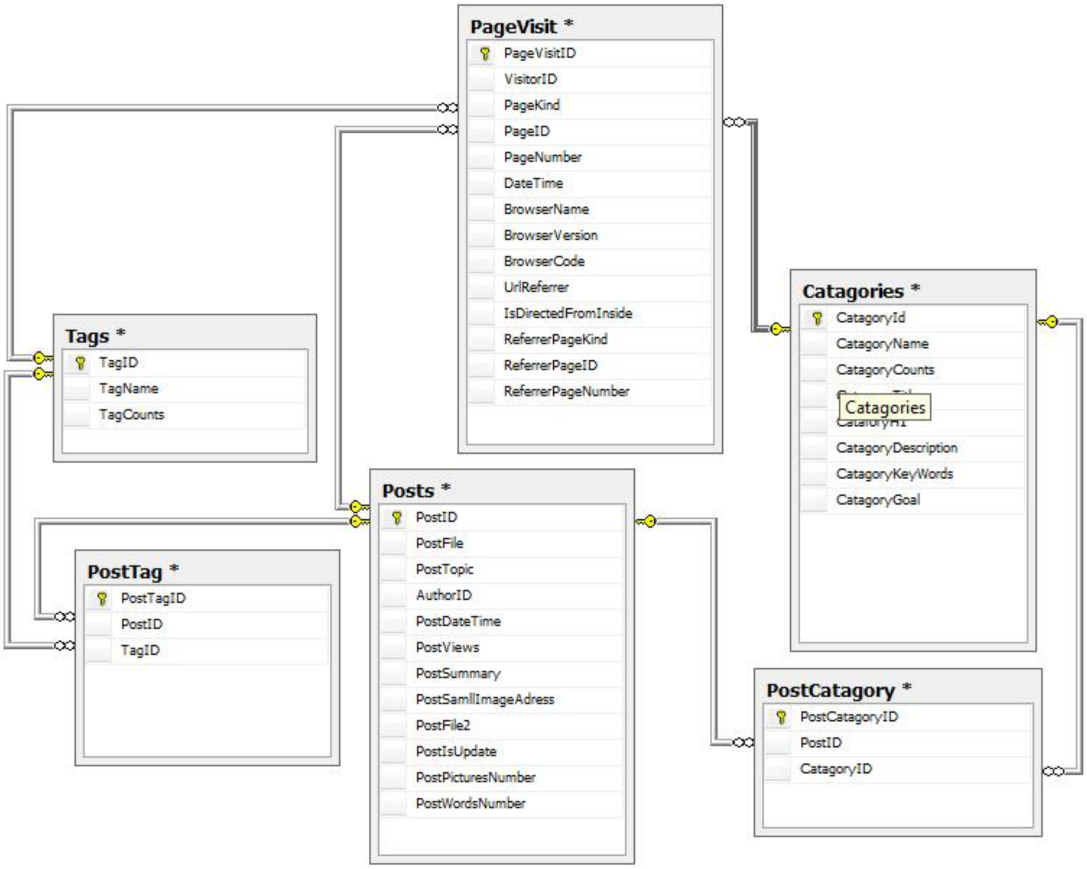


Figure 4 a part of the website's database tables and their relations

Table 3 contains information about the data we selected out of the Website's database. To meet the end of our endeavour that is to select appropriate data out of the available database, we extracted 4929 attributes (columns) for each posts (records). As it was mentioned, the entities we strive to cluster in this paper are pages, so it must not be forgotten that each record owns 4929 attributes. In this table, the column Query is the identification of the SQL query we used to extract that data from the SQL database. These queries, however, are provided in the appendix 1. The last column is the data kind whose values specify the category this data falls under - the three types we mentioned earlier are Content, Structure and Usage data. However, for some reason the content data are divided into first and second group. The first (Content-I) are the kind of data which may have some influence on whether a visitor will visit the post or not, while there is no way that the data associate with the second (Content-II) could have possibly had any influence on the matter.

Table 3 Selected data out of The Website's Database

Attribute name	Description	Query	Number of Attributes	Data Type	Data Kind
Len(Topic)	The length of topic of the posts (*PostTopic*)	Q1	1	Numerical	Content-I
Len(Summary)	The length of the summary of the posts (PostSummary)	Q2	1	Numerical	Content-I
PicID	The ID of Small pictures associated with each post	Q3	1	Nominal	Content-I
NVisits	Total Number of Visits	Q4	1	Numerical	Usage
Popularity	NVisits / Time elapsed since its first day	Q5	1	Numerical	Usage
Cat [1-30]	30 attributes that each is associated with one category in the website. Each attribute shows whether each post belongs to the category or not.	Q6*	30	Categorical (0-1)	Structure
TagN	The total number of Tag pages which are associated with each post	Q7	1	Numerical	Structure
TagPageN	The total number of pages which are associated with the post via the Tags	Q8	1	Numerical	Structure
PostTag [1-1040]	1040 attributes that each stands for each post. The values show how many tags are associating this post (Attribute) to that post (Record).	Q9*	1012	Numerical	Structure
PostVisit [1-1040]	1040 attributes that each stands for each post. The values show how many visitors have visits the pair of pages (Attribute, Record)	Q10*	1012	Numerical	Usage
PostPictureN	The total number of pictures which are used for each post	-	1	Numerical	Content-II
PostWordN	The total number of Words that are used for each post	-	1	Numerical	Content-II
TopicWord [1-2810]	2810 attributes that each one stands for a word. The values show whether the word was used in the post topic or not.	-	2810	Categorical (0-1)	Content-I

It should be noted that some of rows in Table 3 are the representation of only one attribute, whereas the other are the representation of a set of attributes, such as PostTag and PostVisit. These two sets are more like two similarity matrixes which represent similarity of each post with regard to respectively Tags and Visits. In addition, the TopicWord set is a huge pile of data which needs to be reduced and transformed. Firstly, there are nearly 2000 words in the set which have only been used in one post. This kind of attribute contains no interesting information in the sense of clustering task. They removal won't depreciate the extracted information and knowledge but will save us an enormous amount of time analyzing them in no avail. Secondly, in this set of attribute there are lots of prepositions and such like which yet again will not contain any meaningful information inside. Thirdly, to bring the knowledge and information to the surface the remaining of the data in the set were used to construct another similarity measures. To illustrate, the set PostTopicWord, which will be used instead of TopicWord, containing as many columns as the number of the posts, represents how many communal words each pair of posts share in their topics.

4.3 SOM and K-Means comparison

The question whether SOM is an appropriate technique to be used for webpage clustering task can be answered by its frequent usage in the literature. However, in this part we have conducted an experiment showing the superiority of SOM. To do so we compared the consistency of SOM and K-Means with one another. The logic behind is that if an algorithm succeeded in clustering a same data set similarly for several time, it shows that the algorithm is actually successful in finding patterns and the results are not random.

Table 4 the value of SOM's and K-Means's similarity indexes in different experiments

		1th - 2th		**1th - 3th**		**1th - 4th**		**2th - 3th**		**2th - 4th**		**3th - 4th**	
		FM	R	FM	R	FM	R	FM	R	FM	R	FM	R
Content	SOM	0.9607	0.9938	0.9958	0.9994	0.9399	0.9908	0.9625	0.9941	0.9065	0.9852	0.9401	0.9908
	K-means	0.7298	0.9561	0.7803	0.9649	0.7186	0.9518	0.9069	0.9858	0.8482	0.9749	0.8169	0.9701
Structure	SOM	0.8315	0.9225	0.8667	0.9375	0.7435	0.8818	0.9354	0.9725	0.7612	0.9003	0.7461	0.8911
	K-means	0.6922	0.8627	0.7465	0.8830	0.7652	0.8901	0.7966	0.9236	0.8106	0.9265	0.8559	0.9418
Usage	SOM	0.8728	0.9747	0.8791	0.9760	0.8675	0.9737	0.8546	0.9719	0.9908	0.9982	0.8601	0.9730
	K-means	0.5690	0.9273	0.5075	0.9111	0.6626	0.9432	0.6215	0.9294	0.7028	0.9481	0.6682	0.9382

Table 5 Average and the standard deviation of SOM and K-Means and their differences

		Average		**SD**		**Average Difference (SOM – K-Means)**		**SD Difference (K-Means – SOM)**	
		FM	R	FM	R	FM	R	FM	R
Content	SOM	0.9509	0.9924	0.0299	0.0047	0.1508	0.0251	0.0422	0.0078
	K-means	0.8001	0.9673	0.0721	0.0125				
Structure	SOM	0.8141	0.9176	0.0777	0.0338	0.0363	0.013	-0.0211	-0.0033
	K-means	0.7778	0.9046	0.0566	0.0305				
Usage	SOM	0.8875	0.9779	0.0514	0.01	0.2656	0.045	0.021	0.0033
	K-means	0.6219	0.9329	0.0724	0.0133				
Average						0.1509	0.0277	0.0140	0.0026

To that end we clustered each content, structure and usage data four times with both of the algorithms. It means that we clustered content data 4 times using SOM algorithm and 4 times using K-Means algorithm. Structure and usage data also went through the same process. Using Fowlkes–Mallows and Rand indexes [18] the similarities between each four same clusters was calculated. Both of the indexes will have the minimum value of zero and maximum value of one. Whereas the value one for both depicts a complete similarity, the value zero is a sign of thorough discrepancy. The results of the two indexes are presented in Table 4. We kept the algorithm behavioral variable for the both algorithm steady. SOM used a 4x4 hexagonal topology, while the number of clusters for K-means was equivalently 16. For instance, in the table below the two similarity values, Fowlkes–Mallows and Rand indexes, for the 1th and 2th clusters made by SOM using Content data is 0.9607 and 0.9938. Looking briefly at the table, one can conclude that SOM has been developed more similar and consistent clusters. This means that, SOM has been able to defeat randomization and extract more consistent patterns comparing to K-means. Thus, it can be concluded that SOM is superior to K-Means with regard to extracting pattern. However, the extent of its advantage is not discernible from this table. The average difference between SOM and K-Means and also the variance difference between them is shown in Table 5. This table also represents the standard deviation of the two techniques and also their differences. On balance, SOM's FM and R indexes surpasses that of K-Means respectively by 0.1509 and 0.0277. Also, in general, the standard deviation of FM and R Values for SOM is less than K-Means.

4.4 Yardsticks (ground information)

The validation and the test process of a clustering task are not as straightforward as it is for classification or prediction process and for most of cases it requires ground information about the data: some pre-categorized data [17]. Incidentally, there are techniques only to evaluate the consistency of a clustering algorithm. These techniques are not able to give away any information about how well an algorithm has performed. For instance, by randomly bisecting the training data and evaluating the similarity of the clusters, one of them can help the data miner to reach a conclusion whether a technique is consistent or not. No one can deny that without outside information there will be no way to decide whether a clustering technique is doing a good job or not. And to that end, one needs to employ the other technique and try to gather some ground data in order to prove the efficiency of an algorithm.

In this paper we managed to come up with two sets of different ground information. Firstly, there are 30 categories in the dataset and each post is associated with at least one of them. In order to have one column as a ground rule we applied these 30 categories into a sub clustering task using the SOM itself. We performed SOM with 4×4 hexagonal topology which resulted in the first column as a ground rule. Second, in order to be able to engage the visitors' preferences in our ground information we set up a questionnaire by which we asked several of the visitors to give their opinion about each posts. They had to scale every post from 1 to 5 for different types of factors: interestingness, picturesqueness, instructiveness, informativeness, spirituality, and handiness. To illustrate, they were asked to scale, for example, the instructiveness of a post from 1 to 5 based on the impression they would get out of the post title. The reason behind this act is due to the nature of our usage data which is constructed solely on whether users visited a post or not, and other data such as whether they liked it or not, or how long their visit last were not available. However one can see and understand how questionnaire looked like as well as how each factor is defined by referring to the appendix 2.

We managed to have 23 individuals take part in this part of our study and some of whom only filled out questionnaire partly. That is to say, they only scaled the posts based on the first and general factor – interestingness. Although the number of people who have participated seems to some extent insufficient, it must not be forgotten that filling the questionnaire was time-consuming and painstakingly boring. The data extracted from the questionnaires were directly used to construct the second ground information column. 11 full-filled participant's six factor and 11 partly-filled participant's one factor constitute 77 attributes for each post which were used as the input for a 4×4 hexagonal topology SOM to result in the desired column - second column as yardstick.

4.5. Comparison

In order to spot which combination of the data would fit the best to our two ground information clusters we applied 4×4 hexagonal SOM and 4×4 quadrilateral SOM to each and every combination of our data type: Content, Structure, and Usage. However, because we suspected that the combination of Content I and Usage will lead to a better fit we include that as well. So as to differentiate and compare different combinations in the basis that how fit each one would be comparing to ground information columns, we used two different indexes, namely Fowlkes–Mallows index and Rand index [18]. Both of the indexes will have the minimum value of zero and maximum value of one if the clusters are identical. Table 6 present these two indexes' value for all of the data combination. In this table C, S and U respectively stand for Content, Structure and Usage.

Table 6 The values of indexes for different clusters and ground column clusters

		Ground Truth 1 (Categories)		Ground Truth 2 (Questionnaire)	
		FM	R	FM	R
Content	**hex**	0.264	0.892	0.112	0.865
Content	**quad**	0.269	0.887	0.108	0.871
Structure	**hex**	0.215	0.798	0.137	0.747
Structure	**quad**	0.214	0.788	0.135	0.754
Usage	**hex**	0.151	0.87	0.583	0.944
Usage	**quad**	0.163	0.872	0.516	0.935
CS	**hex**	**0.326**	**0.921**	0.106	0.863
CS	**quad**	**0.333**	**0.905**	0.118	0.863
CU	**hex**	0.165	0.867	0.364	0.91
CU	**quad**	0.165	0.87	0.358	0.911
C1U	**hex**	0.157	0.872	**0.712**	**0.961**
C1U	**quad**	0.151	0.873	**0.636**	**0.952**
US	**hex**	0.132	0.852	0.112	0.864
US	**quad**	0.157	0.871	0.111	0.889
CUS	**hex**	0.189	0.881	0.114	0.885
CUS	**quad**	0.187	0.882	0.113	0.886

As we had suspected, one can see the ground truth 2 (questionnaire) were fitted the best with the cluster constructed by the combination of Content-I and Usage data. Nevertheless, the contribution of usage data itself to match is much greater – its index's values are 0.5 and 0.9 which are among the highest. Moreover, it seems that the combination of Content and structure data is fitted significantly better for categories. Although the differences between the other values for both indexes are not as outstanding as the difference for the questionnaire, the gaps seems to be wide enough to conclude the best combination is Content and Structure.

Another point which we tested using a paired-sample t-Test is whether there is significant difference between hexagonal or quadrilateral topology approaching webpage clustering. The assumption that the two topology has no significant influence on the performance of the task could not be rejected (α=0.05). The outcome of the test can be seen in the Table 7. Therefore, it is logical to say that for this website, the type of topology used for SOM clustering should not be seen as an imposing factor.

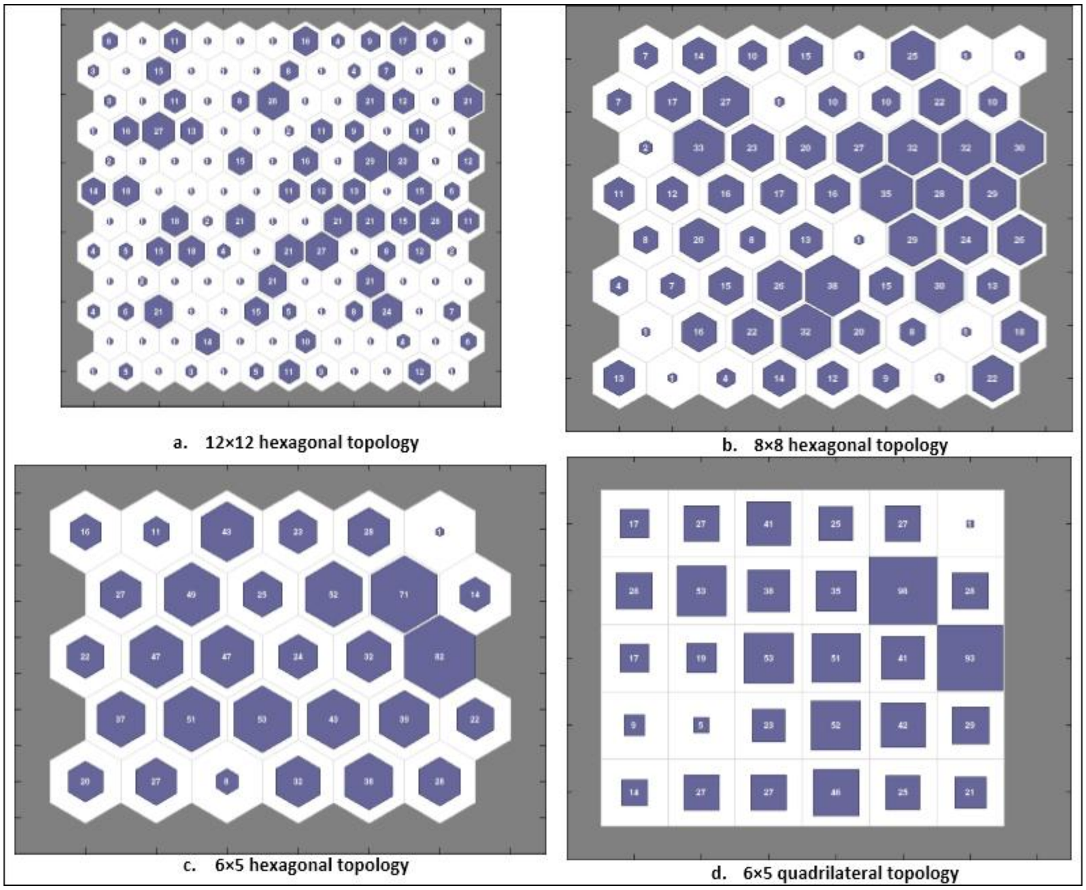
Table 7 t-Test: Paired Two Sample for Quadrilateral and hexagonal topologies

	hex	Quad
Mean	0.55721875	0.55446875
Variance	0.120545015	0.119211225
Observations	32	32
Pearson Correlation	0.998301341	
Hypothesized Mean Difference	0	
df	31	
t Stat	0.767368232	
P(T<=t) one-tail	0.224334056	
t Critical one-tail	1.695518783	

4.6. Webpage profiling

Based on the conclusion we arrived at, there is no doubt over the fact that the combination of Content and structure is the best for the task of webpage profiling. Figure 5 represents different topologies used in order to cluster posts with Content and structure data using SOM algorithm. Part a. of the figure shows the outcome of a 12×12 hexagonal SOM which is obviously too vast a topology to be used for analysis. Topology of part b. hasn't been able to tightly make the clusters so as make us able to analyze different clusters. However, both c and d topologies had been able to effectively distribute the posts in the area. Since we understood from previous that using hexagonal or quadrilateral doesn't significantly affect the clusters outcome, which by the way is fortified by the fact that the two topologies look somehow alike, we opt for 5×6 hexagonal outcome for further analysis.

Using consultation of website owners and through collaboration we analyzed the outcome of clustering task. We used different techniques such as finding the median record, checking the similarities between the members of clusters, and evaluating the differences between the members of other clusters to characterize each cluster. The result is presented in the Figure 6 which is, as a picture, an analogy of the topology used to come up with the clusters. In this figure each cell, each hexagon, has three values inside – two numbers and a string as the cluster description. The left-hand number is the cluster identity starting from one to 30. The right-hand number is the number of records composing of the cluster. It should be noted that the right-hand number is the same as the numbers presented in part c of Figure 5.



a. 12×12 hexagonal topology

b. 8×8 hexagonal topology

c. 6×5 hexagonal topology

d. 6×5 quadrilateral topology

Figure 5 SOM different constructed topologies with Structure and Content

About Figure 6, there is more to the descriptions assigned for each cluster. One can see that there are rather discernable connections between neighbor hexagons. For instance, one can see cells 26, 27, 28, 29, 30, 19, 21, 22, 23, 24, and 17 which are all neighbors have the same general theme – pictures. Interestingly, the cells 21 and 22, which are labeled as news through pictures, are the neighbors of cells 20 and 15, which are generally about news. One can find other such connections between the different cells in the figure and that shows the eligibility and appropriateness of the SOM and the appropriate selection set of data approaching webpage clustering. However, it must not be forgotten that hard as we tried we failed to find a general description for the 16 records composing cluster 25. Apart from the fact that we had to label the cluster as unknown, it is exciting to see that the group of records which could have not been in any group have been stuck in an extreme corner.

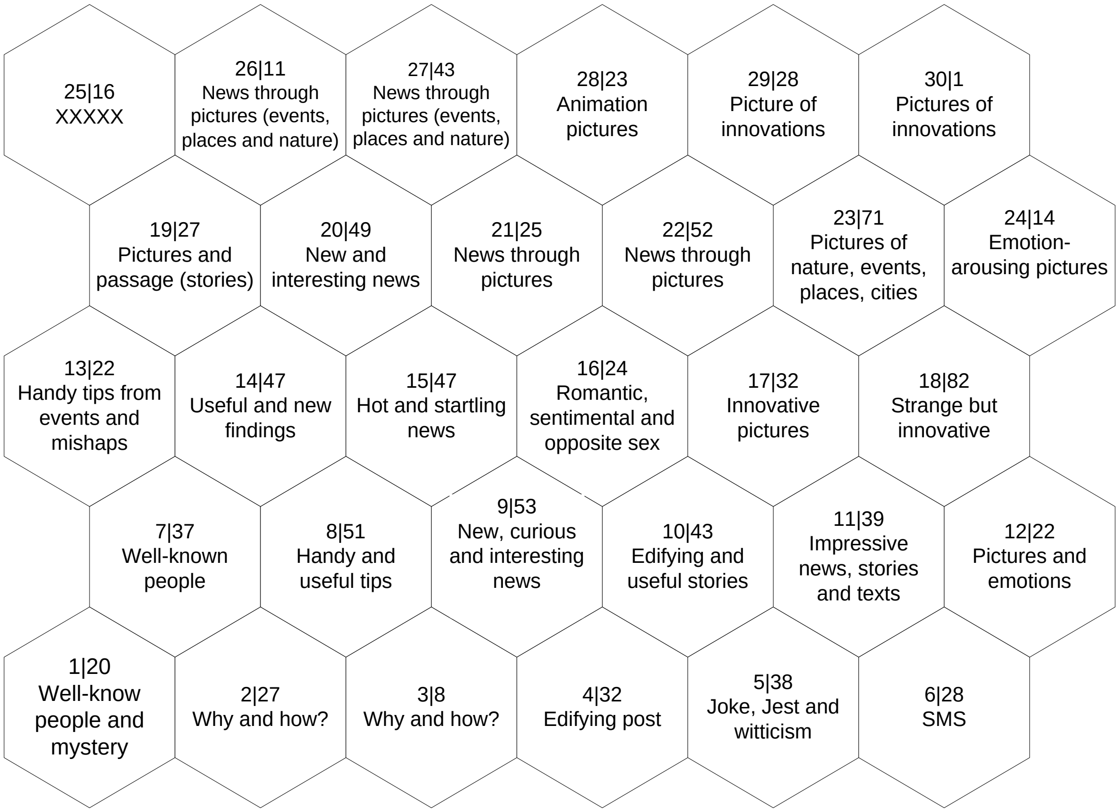


Figure 6 the conceptual topology of clustering by Content and structure Data

4.7. Webpage recommendation using webpage clustering

Websites may have a recommendation system by only using the results of a webpage clustering task based on the visitor’s preferences. Although it’s customarily the job of association rules analysis, we developed a recommendation system by clustering algorithm. The recommendation system works solely on the fact that upon seeing a particular page other members of the page’s cluster are likely to be liked by its visitors. Therefore, in each page the website suggests some of the other members of the page’s cluster as recommendations. Luckily, we actually came to test the effectiveness and influence of such recommendation system on the website. After contriving a cluster based on content and visitors’ behavioural data, we used the result for a recombination system as explained. The influence and the effect of having such system on the website is also observed and analysed.

It was proven that for this website it is best to use content and usage data for extracting visitors’ preference. The clustering analysis was performed using the same data. First, to observe a map of the data we used a rather big hexagonal topology, 12×12 (Figure 7). One can see in the figure that SOM has parted the data in 16 parts. That is to say, SOM is clearly saying the number of clusters needed for this task is 16. However, so as to make our further analysis easier we used 4×4 hexagonal topology and the cluster we finally put into use for the website is presented in Figure 8.

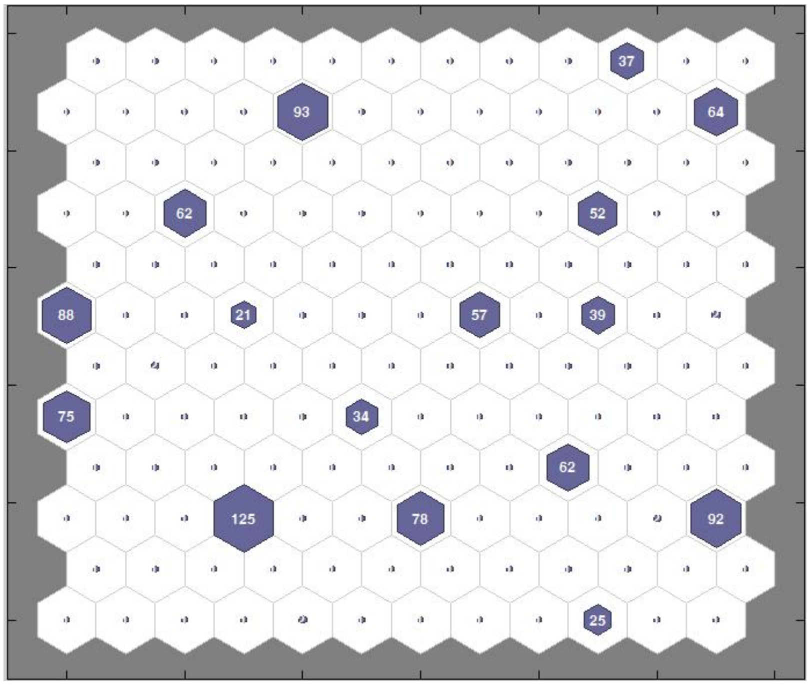


Figure 7 a 12×12 hexagonal SOM constructed topology using Content I and Usage dat**a**

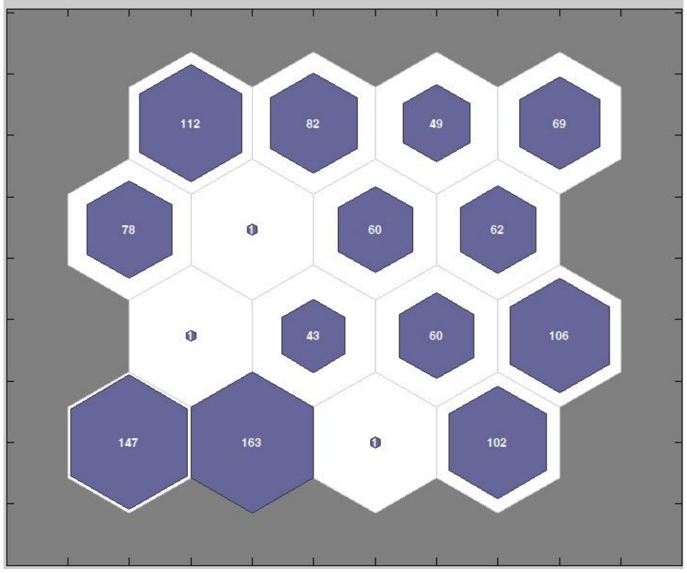


Figure 8 4×4 hexagonal SOM constructed topology using Content I and Usage data

However, in order to make certain that the results of SOM is not random and the algorithm is consistent and is actually extracting pattern, four 4×4 hexagonal were constructed and were compared with one another using Fowlkes–Mallows and Rand indexes. The results of the indexes are presented in Table 8 and the other figures for every other three SOM experiments are shown in appendix 3. The values of indexes in the Figure indicate that the level of similarities between all of the four clusters is high and any of them could be used for the recommendation system. However, we used the cluster 3 having the highest similarity indexes.

Table 8 the comparison of the four 4×4 hexagonal SOM

	C1		C2		C3		C4		Average		R+FM
	R	FM	R	FM	R	FM	R	FM	R	FM	
C1	1	1	0.9505	0.7099	0.9762	0.8632	0.9560	0.7422	0.9707	0.8288	1.7995
C2	0.9505	0.7099	1	1	0.9570	0.7526	0.9477	0.6926	0.9638	0.7888	1.7526
C3	0.9762	0.8632	0.9570	0.7526	1	1	0.9619	0.7803	0.9738	0.849	1.8228
C4	0.9560	0.7422	0.9477	0.6926	0.7803	0.9619	1	1	0.921	0.8492	1.7702

Table 9 eleven samples of days for both with and without recommendation feature (ER, APPV and ADV indexes)

15th Nov 2013 – 28th Dec 2013 (WITHOUT)					29th Dec 2013 – 12th Feb 2014 (WITH)				
N.	Date	ER (percent)	APPV (page)	ADV (second)	N.	Date	ER (percent)	APPV (page)	ADV (second)
1	2013 19th Nov	51	2.64	837.28	1	2013 31th Dec	50	2.84	828.55
2	2013 21th Nov	56	2.5	749.87	2	2014 1th Jan	56	2.57	846.85
3	2013 27th Nov	60	2.33	620.45	3	2014 5th Jan	52	2.66	841.12
4	2013 30th Nov	59	2.42	645.87	4	2014 8th Jan	54	2.62	647.82
5	2013 2th Dec	60	2.43	729.34	5	2014 12th Jan	55	2.59	693.1
6	2013 11th Dec	64	2.23	558.59	6	2014 20th Jan	55	2.41	791.43
7	2013 12th Dec	62	1.92	524.98	7	2014 21th Jan	61	2.36	607.4
8	2013 17th Dec	59	2.52	772.14	8	2014 26th Jan	57	2.64	756.66
9	2013 21th Dec	59	2.32	548.70	9	2014 4th Feb	56	2.55	807.87
10	2013 25th Dec	59	2.36	611.79	10	2014 5th Feb	57	2.56	763.76
11	2013 26th Dec	63	2.36	598.433	11	2014 8th Feb	59	2.57	703.33

From 29th Jan 2013 a new feature was added to Khatekhalagh website. The feature was up and working until 12th Feb 2014 which rounds it up to approximately a month and 12 days. In every post page visitors were exposed to a new feature named “Our Suggestion”. In this part, as delineated before, 15 other posts which are in the same clusters with the page are suggested to the visitor. These 15 posts can be any of the other members of the cluster with different probability. That is to say, a more popular post in general has more chance to be suggested to visitors. However, popularity in general is calculated by dividing the number of a page visits by the number of days the post’s been exposed in the website.

To observe the influence of having this new feature in a website we compared and contrasted some successfulness indexes namely exit rate (ER), the average number of pages per visits (APPV), and the average duration of the visits (ADV). The feature was put on website for one month and twelve days. As to compare its influence we compared it to the data from the same amount of time before its having been installed. Since the task of extracting the aforementioned successful indexes were time consuming and also we needed to defeat the influence of others factors involvement, 11 random days were selected from each period to be used for comparison. These two sets of 11 days and the three indexes' values for them are presented in Table 9. Moreover, so as to test whether the samples are from a normal distribution we used Shapiro Wilks's normality test – n is 11 which is between 3 and 2000. Also to test the equality of Variance, Levine's Test was employed. The results are presented in Table 10 and it shows that the assumptions of all of the indexes sample following a normal distribution and the three pairs having equal variance cannot be rejected (p-v>0.05). To conclude, looking at the Table 11 which is the results of three different t-tests on the assumption that the indexes' value after having installed the recommendation feature remains the same, one can see that for all of the indexes the assumption of equality has been rejected (p-v<0.05) and this, in turn, prove the point that by having installed this new feature the website's successfulness indexes actually had taken a turn for the better.

Table 10 Shapiro Wilks normality test and Levene's Test for Equality of Variance

	Shapiro Wilks normality test			Equal Variance	
	df	**Stat**	**p-value**	**Stat**	**p-value**
With-ER	11	0.923	0.346	0.007	0.933
Without-ER	11	0.886	0.126		
With-APPV	11	0.923	0.346	0.699	0.413
Without-APPV	11	0.905	0.214		
With-ADV	11	0.929	0.465	1.032	0.322
Without-ADV	11	0.932	0.434		

Table 11 t-student test results

Pairs	**df**	**Stat**	**p-value**
ER (With and without)	20	2.591	0.017
APPV (With and Without)	20	-3.147	0.005
ADV (With and Without)	20	-2.518	0.02

5 Conclusion and future trend

In this paper we employed SOM and K-Means to tackle webpage clustering task for an Iranian website so as to answer some questions and doubts we had come across in the literature. Since we had seen that webpage clustering is conventionally done using one of the three types of data – Content, Structure, and Usage - we were interested to evaluate the different role of each type on the outcome of the task. We came to the conclusion that for the website used in this study the combination content and usage data can best extract the visitors' preferences and also the combination of structure and content will do best for webpage profiling task. This fact, in turns, shows the inappropriateness of the common belief that content data is only appropriate for webpage profiling and usage data should be used for visitors' preference. Although using the outcome of this study we cannot assert that, for instance, for webpage profiling one has to use amalgamation of content and structure data, the necessity of taking the zero step so as to pick and choose the right set of data has been shown. Also, using the data of the website our suspect about the superiority of SOM over K-Means in dealing with webpage clustering task were proved. Moreover, it was tested and proved that being hexagonal or quadrilateral has no influence on the performance of SOM in dealing with webpage clustering.

Another exciting part of the study was the use of the result and insight gained from the experiments. Based on them, we conducted two webpage clustering task: one to profile and characterize the webpages in the website and the other to put together the pages similar with regard to visitor's preferences. About the webpage profiling, although we didn't manage to compare or evaluate quantitatively the result of using the clustering upshot, the outcome proved appropriate and sufficient by being reasonable and arguable. And second, about the clustering for a recommendation system, we came to have the proposed system installed on the website. The successfulness indexes of website when comparing showed that the website on the whole had improved because of the new feature. The three indexes were meaningfully and significantly different which showed the positive influence of the new feature on the website.

In the course of this study we tried to follow the methodology of data mining, first to make our work valid and easier to follow, and second to avoid misjudgements. This fact empowered us to be able to show the whole process we went through form the selection of data to the assessment of the results. As we discussed, although the Iranian web environment has already seen fierce competitions, there is little evidence about the current usage of data mining in this country's websites. This actually is another reason that we used the methodology of data mining and that is to pave the way for Iranian websites to be able to incorporate these state of the art series of technologies into their design.

5.1. Limitations

We had to deal with different complications through the study. First, as hard we searched we were not able to find a very well-known website willing to help and provide us with the data we needed. Second, as it was mentioned we were not provided with the full access to the website under the study so we reduced to exploit as much as we could. Last but not least, a data mining process is not completed unless we test our finding in real world. Because this study was not seen as practical task by the website owner, we only managed to have one part of our study tested and analysed online and we failed to have the data mining process completed for webpage profiling.

5.2. Applicability and Future Trends

Our findings are to be used by both researcher and website owner. Form a researcher point of view there are several directions. As we tested the new recommendation feature on the case, one may want to have this paper finding tested in other real websites. In addition, as we discussed by the data we had at our disposal we only managed to draw attention to the need of taking a zero step so as to find the proper set of data and we found out that only for our case content and usage data are appropriate for webpage recommendation system and the combination content and structure can better the result of webpage profiling. Therefore, testing the consistency of these finding on other websites can be the subject of another study. From a website owner's point of view, this study also has several values. Among others, we depicted a whole data mining process form the selection of data to having our result tested logically and statistically. Right from the word go their website database design might be compared to this paper's database to see whether their website is ready to have data mining techniques set in their website. However, this is only a repetition to mention, the only way that a data mining technique can operate is with the presence of appropriate data. Moreover, similar websites such as news agencies might want to use the depicted approach in this paper to get to know their websites better and make their website ready for further improvements.

References

[1] S.-T. Yuan, H.-S. Chen, A study on VRM-awareness enterprise websites, Expert Systems with Applications, 22 (2002) 147-162.

[2] S. Park, N.C. Suresh, B.-K. Jeong, Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm, Data & Knowledge Engineering, 65 (2008) 512-543.

[3] M.J.A. Berry, The Virtuous Cycle of Data Mining, in: Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, Wiley, Indiana, 2004.

[4] C.-C. Lin, L.-C. Tseng, Website reorganization using an ant colony system, Expert Systems with Applications, 37 (2010) 7598-7605.

[5] K.A. Smith, A. Ng, Web page clustering using a self-organizing map of user navigation patterns, Decision Support Systems, 35 (2003) 245-256.

[6] B. Prasetyo, I. Pramudiono, K. Takahashi, M. Kitsuregawa, Naviz:Website Navigational Behavior Visualizer, in: M.-S. Chen, P. Yu, B. Liu (Eds.) Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, 2002, pp. 276-289.

[7] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive document collection, Neural Networks, IEEE Transactions on, 11 (2000) 574-585.

[8] S.-H. Huang, H.-R. Ke, W.-P. Yang, Structure clustering for Chinese patent documents, Expert Systems with Applications, 34 (2008) 2290-2297.

[9] Z. Su, Q. Yang, H. Zhang, X. Xu, Y.-H. Hu, S. Ma, Correlation-based web document clustering for adaptive web interface design, Knowledge and Information Systems, 4 (2002) 151-167.

[10] A. Ypma, E. Ypma, T. Heskes, Categorization of Web Pages and User Clustering with Mixtures of Hidden Markov models, Proceedings of the International Workshop on Web Knowledge Discovery and Data Mining, Edmonton, Canada, (2002) 31--43.

[11] D. Qi, C.-c. Li, Self-Organizing Map based Web Pages Clustering using Web Logs.

[12] Y. Kim, Weighted order-dependent clustering and visualization of web navigation patterns, Decision Support Systems, 43 (2007) 1630-1645.

[13] G.E. Tsekouras, C. Anagnostopoulos, D. Gavalas, E. Dafni, Classification of Web Documents using Fuzzy Logic Categorical Data Clustering, International Federation for Information Processing, 247 (2007) 93-100.

[14] T. Kohonen, Self-organizing maps, Springer, 2001.

[15] M.J. Berry, G.S. Linoff, Artificial Neural Networks, in: Data mining techniques: for marketing, sales, and customer relationship management, Wiley. com, 2004.

[16] P.-N. Tan, Introduction to data mining, Pearson Education India, 2007.

[17] D. Delling, M. Gaertler, R. Görke, Z. Nikoloski, D. Wagner, How to evaluate clustering techniques, Univ., Fak. für Informatik, Bibliothek, 2006.

[18] M. Meilă, Comparing clusterings—an information based distance, Journal of Multivariate Analysis, 98 (2007) 873-895.

Appendix 1 – SQL Queries

Q1

```
select LEN(PostTopic) from Posts
```

Q2

```
select LEN(PostSummary) from Posts
```

Q3

```
select PostID,kkk.PictureID from Posts inner join(
  select kk.PostSamllImageAdress,Count, ROW_NUMBER() OVER (ORDER BY Count desc) as
PictureID from
  (select PostID, Posts.PostSamllImageAdress,count from Posts inner join
  (select PostSamllImageAdress, COUNT(PostID) as count
  from Posts
  group by PostSamllImageAdress) as k on Posts.PostSamllImageAdress =
k.PostSamllImageAdress)as kk
  group by kk.PostSamllImageAdress,Count
  )
  as kkk on kkk.PostSamllImageAdress = Posts.PostSamllImageAdress
  order by PostID
```

Q4

```
select Posts.PostID, COUNT(PageVisitID) from Posts inner join (select * from PageVisit
where PageKind='Post')k
                      on Posts.PostID = k.PageID
group by Posts.PostID
order by Posts.PostID
```

Q5

```
select PostID, sort from (
       SELECT        TOP (2000) PostID, PostFile, PostTopic, AuthorID, PostDateTime,
PostViews, PostViews / CONVERT(float, GETDATE() - PostDateTime) AS sort
       FROM                dbo.Posts
       where PostDateTime <  (select MAX(DateTime) from PageVisit)
       ORDER BY sort DESC, LEN(PostTopic) DESC ) as kk
       order by PostID
```

Q6

```
select PostID,CatagoryID from PostCatagory
```

Q7

```
select Posts.PostID, COUNT(TagID) from Posts inner join PostTag
                              on Posts.PostID = PostTag.PostID
               group by Posts.PostID
               order by Posts.PostID
```

Q8

```
select TagID, count(PostID) from PostTag
       group by TagID
       order by TagID
```

Q9

```
select postID from PostTag
       where TagID in (
       select TagID from PostTag
       where postID='320')
```

Q10

```
select VisitorID, PageIDs = stuff((select ','+ CONVERT(varchar,PageID) from PageVisit
b where PageKind='Post' and VisitorID=a.VisitorID and b.VisitorID = a.VisitorID FOR
XML PATH('')), 1, 2, '' ) from PageVisit a
group by VisitorID, PageKind
having PageKind='Post' and COUNT(PageVisitID)>1
order by VisitorID
```

Appendix 2 – The Questionnaire

Khatekhalagh is an entertaining website. It contains lots of different information, story, news, advices in various areas which are presented in distinct pages that we call posts. When visitors come to this website they must decide whether they would like to see a post solely based on the topic they read. Here there are around 1000 of the website's posts that we need you to scale from 1 to 5 based on how you feel the topic appeals to you according to the different factors. These factors and their definition are presented below:

Interestingness: Based on the topic how interesting do you think the post would be for you. Or if you were visiting the website how much you would want to see this post.

Picturesqueness: Based on the topic do you think the post would appeal to you due to the probable pictures inside it. In another word, you are to give 1 to the post if the topic gives you the impression that the post does not contain pictures and give 5 to the post if you are under impression that the post contains pictures which would made you want to see the post.

Instructiveness: Based on the topic do you think the post would appeal to you because it'll add something useful to you. You are to give 1 to the post if you don't think the post would help you in anyway if you were to see it and give 5 to the post if you would actually want to see the post only because you feel that it would add something valuable to you.

Informativeness: Based on the topic do you think the post would appeal to you because it can inform you about something you didn't know before. You are to give 1 to the post if you don't think there is anything new for you and give 5 to it if you actually want to see the post because you feel there is something in the post you would like to know.

Spirituality: Based on the topic do you think the post would appeal to you because it would touch a chord with you and you would like to see the post because it would make you feel good. You are to give 1 to the post if there is nothing touchy about the post in your opinion and give 5 to it if you would like to see the post only because your inner spirits wants it.

Handiness: Based on the topic do you think if you read the post it would become handy for you in future. You are to give 1 to the post If you think that there is nothing handy about the post and give 5 to it if you would feel like seeing it only because you would find it handy in future.

Post topic	Interestingness	Picturesqueness	Instructiveness	Informant	Spirituality	Handiness
عظمت آفرینش پروردگار						
هنرنمایی با تراشه های مداد						
گردنبند هایی از موی انسان						
نقاشی اتومبیل ته تو						
حیواناث عظیم الجثه						
چیدمان جالب قوطی ها						
....						

Appendix 3 – The Other Figures

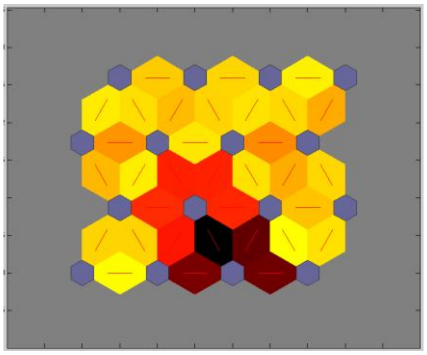


Figure 9 SOM neighbor distances plot – 1th run

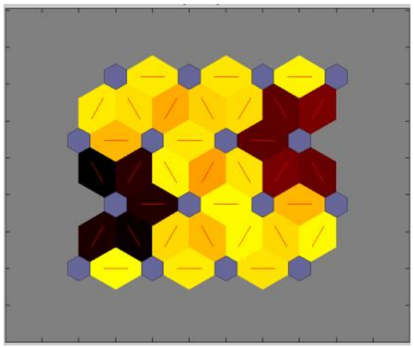


Figure 10 SOM neighbor distances plot – 2th run

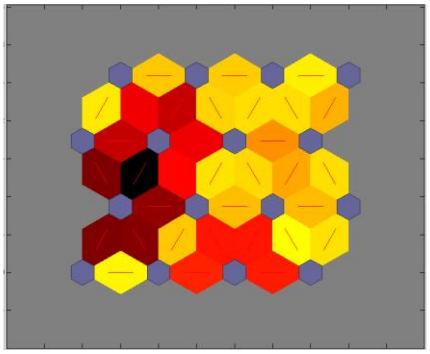


Figure 11 SOM neighbor distances plot – 3th run

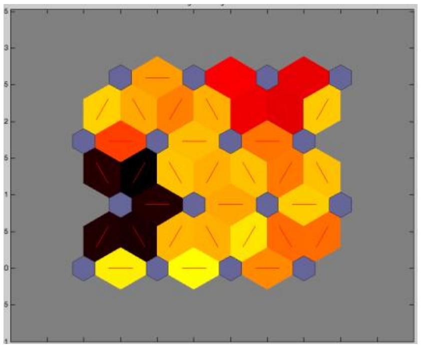


Figure 12 SOM neighbor distances plot – 4th run

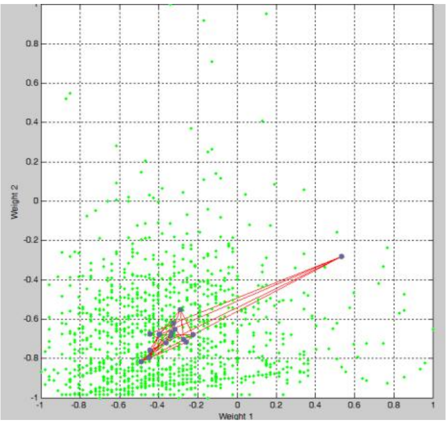


Figure 13 SOM weight positions plot – 1th run

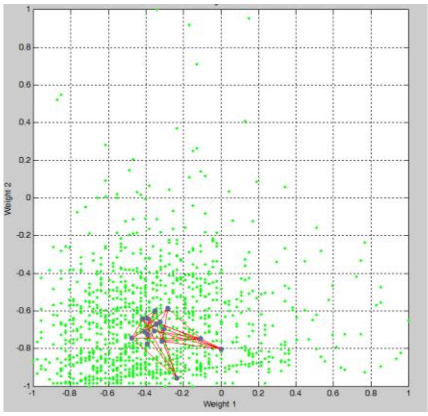


Figure 14 SOM weight positions plot – 2th run

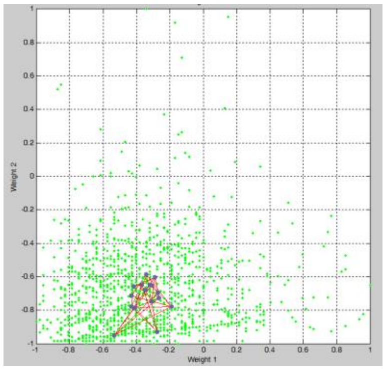


Figure 15 SOM weight positions plot – 3th run

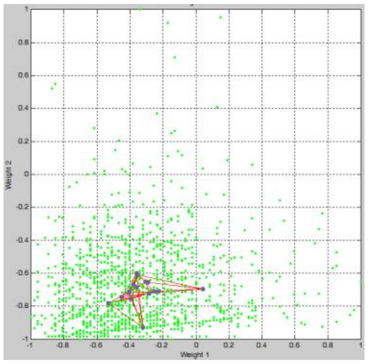


Figure 16 SOM weight positions plot – 4th run

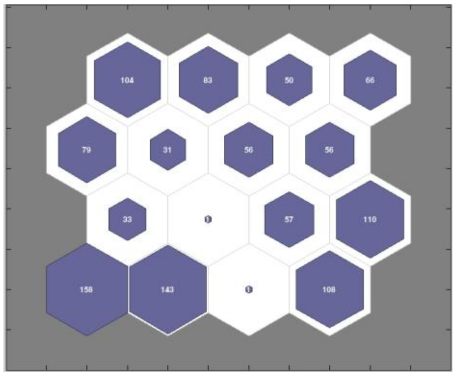


Figure 17 SOM sample hits plot – 1th run

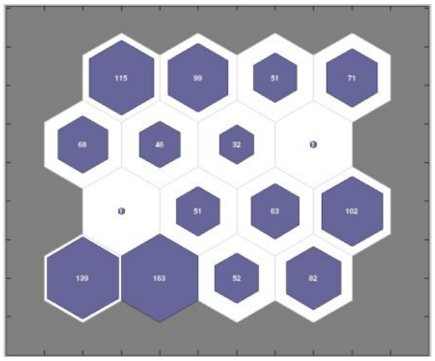


Figure 18 SOM sample hits plot – 2th run

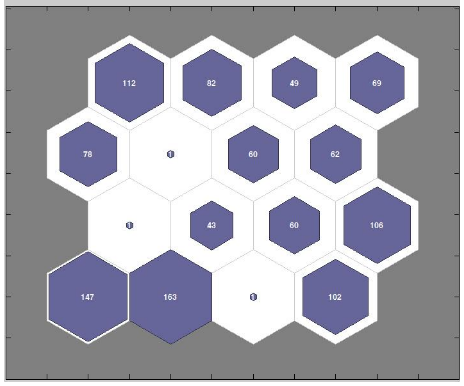


Figure 19 SOM sample hits plot – 3th run

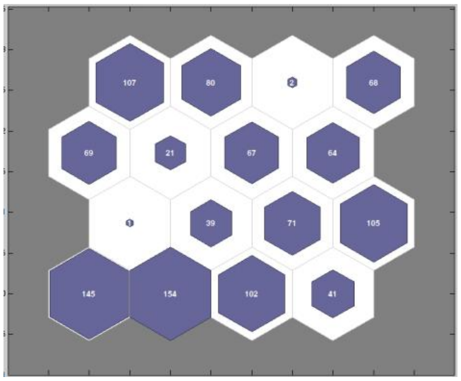


Figure 20 SOM sample hits plot – 4th run