SOMvisua: Data Clustering and Visualization Based on SOM and GHSOM

Mohammad A. Mikki, Khalid M. Kahloot

Abstract— Text in web pages is based on expert opinion of a large number of people including the views of authors. These views are based on cultural or community aspects, which make extracting information from text very difficult. Search in text usually finds text similarities between paragraphs in documents.

This paper proposes a framework for data clustering and visualization called SOMvisua. SOMvisua is based on a graph representation of data input for Self-Organizing Map (SOM) and Growing Hierarchically Self-Organizing Map (GHSOM) algorithms. In SOMvisua, sentences from an input article are represented as graph model instead of vector space model. SOM and GHSOM clustering algorithms construct knowledge from this article.

SOMvisua provides a visual animation for eight famous graph algorithms execution with animation speed control. It also presents six types of visualization. For visualization of similarity lists, we use well-known methods that take a similarity list as input and according to the used similarity measure; an adjustable number of most similar sentences are arranged in visual form. In addition, this paper presents a wide variety of text searching. We conducted experiments on the SOMvisua using a large document dataset. Then we compared the performance with that of hierarchal clustering with automated topology based SOM and GHSOM clustering to prove the superiority of SOMvisua.

Index Terms— Clustering, visualization, Self-Orgainzing Map, Growing Hierarchically Self-Organizing Map, text search, text similarity.

I. INTRODUCTION

Text from distributors or in web pages is based on expert opinion of a large number of people including the views of authors. It has different cultural or community aspects, which make extracting information from it very difficult. Search in text is to find text similarities between the sentences, paragraphs, and articles.

In the past few years, the importance of research in the field of search in textual information has become very important. A lot of research in text information retrieval has been carried out recently with the main concern in extraction and analysis of the text to describe the different aspects of the view of information. This information is contained mainly in three types of sources: First, reference text file from digital documents. Second, metadata provided by the distributors of web sites. Third, information extracted from the Internet. Text analysis is based on the descriptive function at a high level of the context, like the matrix structure presented in [1] or graphical representation based on the text of the attributes of the metadata described in [2]. This describtive content can be clearly found in encyclopedias website such as Wikipedia, Encyclopedia.com, and Webopedia etc.

This paper introduce a framework for data clustering and visualization called SOMvisua. SOMvisua is based on a graph representation of data input for **Self-Organizing Map (SOM)** and **Growing Hierarchically Self-Organizing Map (GHSOM) algorithms**. In SOMvisua, sentences from an input article are represented as graph model instead of vector space model.

SOMvisua supports one of the most important tasks in text information retrieval that is extracting similarities and building a structure for data representation in addition to text similarities graph visualization. Experiments on the SOMvisua show it superiority over SOM and GHSOM. Phase one of SOMvisua is features extraction, which implemented using Google PageRank algorithm.

I.1 Feature Extraction

Feature Extraction is a branch of pattern recognition and image processing. Dimensionality reduction is the main purpose of feature extraction. For large input data set, there will be redundant suspected to an algorithm. Therefore, the input data will be transformed into a reduced representation set of features [3]. The transformation of input data into the set of features is called feature extraction. In this paper, we select Google's PageRank algorithm for feature extraction.

- *Mohammad A. Mikki is with the computer-engineering department, Islamic University of Gaza (IUG), Gaza, Palestine.*
- *Khalid M Kahloot is with the department of Information technology, Ministry of Education, Gaza, Palestine.*

I.2 Google PageRank algorithm

Over the past few years, Google is so far the most widely, used search engine in the world. PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages.

According to Google, PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that websites that are more important are likely to receive more links from other websites.

It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best known. Google uses an automated web spider called Googlebot to actually count links and gather other information on web pages [4].

I.3 Visualization

There are 4 visualization types. These types are:

I.3.1 Circled Bars Visualization:

The Circled Bars visualization approach offers a simple method to answer questions like; "Which sentence produces similar concept to that of a selected sentence A?" It thus takes a similarity list as input. Given a seed sentence A, an adjustable number of most similar sentences (according to the used similarity measure) are arranged in a circle. The sentences are ordered by their similarity to sentence A. Filled arcs that vary in length and color corresponding to the applied color map visualize the similarity values. A Circled Bars visualization is generated from Google PageRank algorithm and the color map Fire is applied. Hence, the values in parentheses after the sentence names indicate the probability for the respective sentence to be found on a webpage that is known to mention the seed sentence A.

I.3.2 Circled Fans Visualization:

The Circled Fans visualization is a conceptual extension of the simple Circled Bars. While the Circled Bars only take the nearest neighbors of a given seed sentence (or any other entity) into account, the Circled Fans incorporates similarities in a transitive manner.

I.3.3 Probabilistic Network Visualization:

A Probabilistic Network visualization is based on a similarity graph of concept sentences. Using this method, first, the vertices representing the data items are placed randomly on the screen. Then, an adaptation process that moves similar data items closer to each other is performed iteratively. Finally, edges between data items are drawn with a probability that is proportional to their similarity. The size of each vertex is calculated using the summed similarities between the data item represented by the vertex and all other data items. The label of a vertex is displayed when the mouse is moved over it.

I.3.4 Sunburst Visualization:

The Sunburst as proposed in is a circular, space-filling visualization technique for illustrating hierarchical data. It is sometimes also referred to as InterRing. The center of the visualization represents the highest element in the hierarchy, whereas arcs further away from the center illustrate elements on deeper levels. Child elements are drawn within the angular borders of their parent, but at a more distant position from the center.

The rest of the paper is organized as follows: Section II presents SOM and GHSOM clustering algorithms. Section III presents related work. Section IV presents SOMvisua. Section V presents SOMvisua experemental results. Finally, Section VI concludes the paper.

II. SOM and GHSOM CLUSTERING ALGORITHMS

II.1 SOM

SOMs are a data visualization technique invented by Professor Teuvo Kohonen, which reduce the dimensions of data using self-organizing neural networks. The problem that data visualization attempts to solve is that humans simply cannot visualize high dimensional data as is so techniques are created to help us understand this high dimensional data. The way SOMs go about reducing dimensions is by producing a map of usually 1 or 2 dimensions which plot the similarities of the data by grouping similar data items together. So SOMs accomplish two things, they reduce dimensions and display similarities [5].

The self-organizing map (SOM, Kohonen-Map) is one of the most prominent artificial neural network models adhering to the unsupervised learning paradigm. The model consists of a number of neural processing elements, i.e. units. Each of the units i is assigned an n-dimensional weight vector

m_i. It is important to note that the weight vectors have the same dimensionality as the input patterns [8].

Since its development, SOM has been used as a powerful tool for data clustering [6]. SOM can analyze data sets with varying statistics such as sizes, shapes, density distribution, overlaps, etc. Moreover, SOM has the capability of data clustering without defining the number of clusters [7].Traditional clustering algorithms such as k-means do not provide such important capabilities, which make SOM better.

II.2 GHSOM

In spite of the stability and popularity of the SOM, at least two limitations have to be noted, which are related, on the one hand, to the static architecture of this model, as well as, on the other hand, to the limited capabilities for the representation of hierarchical relations of the data [8].

We carried out modifications over GHSOM to address both limitations. The modified version is an artificial neural network model with hierarchical architecture composed of independent growing self-organizing maps with graph-based. By providing a global orientation of the independently growing maps in the individual layers of the hierarchy, navigation across branches is facilitated [8].

The key idea of the original GHSOM is to use a hierarchical structure of multiple layers where each layer consists of a number of independent self-organizing maps (SOMs). Once SOM is used at the first layer of the hierarchy. For every unit in this map a SOM might be added to the next layer of the hierarchy. This principle is repeated with the third and any further layers of the GHSOM [9].

III. Related Work

Do Phuc and Mai Xuan Hung have developed a system for clustering the graphs [10]. They use SOM neural network for clustering the graphs and extracting the main ideas from the documents. They make SOM put the documents on a document map and help to access the content of similar documents.

Distance-based similarities in neighborhoods are represented in most of the proposed visualization schemes. Some of those schemes are U-matrix [11] and its variants [12]. The size and shape of the cells to represent the prototypes are described in [13]. Alternatively, some methods use Euclidean distances to update the grid positions of the prototypes for visual inspection as adapted lately in [14], double SOM, and visualization-induced SOM (ViSOM) [15]. Size of receptive fields [13] and smoothed histograms [16] are other methods that use density-based visualizations. However, density-based representations are less helpful compared to the distance-based visualizations unless density representation has a higher resolution than the receptive field size.

Some proposed studies tried to solve the weaknesses of VSM by graph-based models. Most of these studies improved successfully the quality of the resultant clusters. Semantic graph is used for document clustering as in [17]. In this algorithm, a semantic graph is used to represent semantic relationships in documents then convert those graphs into vectors. Vectors are then used in classical SOM algorithm as input. An improvement in the quality of document clustering is shown in this algorithm but it did not propose a direct technique to use the semantic graphs directly with SOM.

H. Chim and X. Deng in [18] proposed Suffix Tree Clustering (STC) algorithm, which is a phrase-based document clustering approach. The time complexity of STC is **nlog(n).** STC produces a high quality resultant clustering.

Several studies that relate to using SOM for generation of topological maps of textual documents have been published. A. Becks, S. Sklorz and M. Jarke [19] use SOM as visualization method, which allows easy access to enterprise document collection. They suggested a modularization of similarity definition.

K. Lagus, T. Honkela, S. Kaski and T. Kohonen [20] developed WebSOM, visualization system for exploration of large collection of Internet Newsgroup e-mails. Documents were mapped from their n-dimensional document content space to two-dimentional map of neurons with topology preservation. After this, each neuron is labeled with Newsgroup name that most documents mapped to this neuron belong to.

L. Soergel and Marchionini [21] use SOM to construct a self-organizing map for information retrieval. They create a map of AI literature documents and later divide map into regions – word areas.

IV. SOMvisua APPROACH

SOMvisua uses SOM and GHSOM for clustering of the text similarity features. SOMvisua also presents similarity graph visualization with the ability to apply eight famous graph algorithms. All of that enables us to finally use the SOMvisua with a wide variation.

SOMvisua is an automated hierarchical growing due to the nature of GHSOM and the hieratical advantage of Google PageRank similarity graph. The framework is able to capture different types of clusters by using appropriate similarity criterion. The number of clusters is determined by prior knowledge on data sets using a recent cluster validity index derived from Google PageRank. SOMvisua shows that a graph input-output clustering can produce better partitioning than other types of clustering.

SOMvisua is based on graph visualization of the topology of the neutral network resulting documents clustering using a modified version of SOM. This method combines the advantages of graph representation in both inputting and outputting to the clustering process. Well illustrated relationships representing the semantic in documents as a graph is the input and illustrated visual representation of topology of the resulting clustering is the output.

SOMvisua provides a file input and output mechanism to load and save data graphs as ASCII-text files. It generates metadata files as ASCII-text files and offers the ability to preserve a complete workspace as a set of data graphs and meta-data. We present a full implementation of a graph input SOM and offer more options for finitialization methods including random, linear, and gradient or Su, Liu and Chang algorithm (SLC) [22]. To cover all potential needs of farther development, we prepare SOM with two training methods. These are sequential and batch job methods. We also enhance the implementation of SOM with functions such as calculate SOM, show SOM-grid optionally, with labels from metadata and load and save SOM-objects from previously saved process. As an extension of SOM, we present an implementation of the GHSOM as well.

Basic structure of SOMvisua is shown in Figure 1. The "set of sentences" is the sentence space wanted to be explored. SOMvisua will generate the similarity Graph of this document set via Google PageRank algorithm. A similarity graph shows sentences as vertices. SOMvisua provides multiple methods for visualizing the similarity graph. SOMvisua provides the capability of executing eight well-known graph algorithms for visualization. The similarity graph is considered as an input to "SOM clustering" algorithm. SOMvisua passes similarity graph to "SOM clustering" algorithm to generate the map. Similar sentences will be placed close to each other on the map. Beside this topological structure, sentences are grouped into clusters. Sentences that are very similar and best described with the same unit map are put in the same cluster.

In SOMvisua, modified SOM is fed with the similarity graph that is generated by text pre-processing stage that is implemented using Google PageRank algorithm. Each sentence is represented with one graph node. This is important because we want clustering at the level of neuron and use SOM only as visualization method. After learning the SOM, each sentence graph is mapped to one of the map neurons that is used for this sentence. Each point represents a sentence and each line crosses a neuron.

SOMvisua provides four major capabilities, which are:

- File input-output
- Data processing
- Feature extraction/Text processing
- Visualization

File input-output provides capabilities to manipulate data files. To facilitate preserving our project's status, we provide tools to load and save data graphs generated by our project provided as ASCII-text files. Other tools to load and save metadata files provided as ASCII-text files are provided. Tools to load and save the complete workspace are provides as well.

Data processing constructs data structures to organize data in SOMvisua. Moreover, it assigns names to the data graphs and metadata lists with the ability to rename data structures. We provide a data graph decomposition into single lists by sources or by destinations. Data processing is also used for normalization of data graphs.

Feature extraction/Text Processing is implemented through Google PageRank algorithm. Google PageRank algorithm provides a graph-based algorithm for text summarization.

Visualization is used by SOMvisua to visualize similarity graph with the ability to apply eight

famous graph algorithms. SOMvisua uses four types of visualization:

- Circled Bars,
- Circled Fans,
- Probabilistic Network and
- Sunburst.

For visualization of similarity Lists; Circled Bars visualization is used.

Circled Bars visualization method is a well-known method that takes a similarity list and given a seed sentence as input. According to the used similarity measure, an adjustable number of most similar sentences are arranged in a circle. The sentences are ordered by their similarity to the seed sentence.

The **Circled Fans visualization** is a conceptual extension of the simple Circled Bars. While the Circled Bars only take the nearest neighbors of a given seed sentence (or any other entity) into account, the Circled Fans incorporates similarities in a transitive manner.

The **Probabilistic Network visualization** for similarity graphs uses a graph-based model for illustrating similarities between graphs by using one prototype for each of a number of given classes of sentences.

Sunburst visualization for term co-occurrences starts with the whole set of the terms with the highest document frequency and are selected and visualized as filled arcs around a centered circle called the root node that represents the entire document collection.

SOMvisua is conducted through the following steps:

1. Phase one of SOMvisua is features extraction where most important features of the sentences are found using Google PageRank algorithm. Visualization of input and output graphs are presented in a later step.
2. The second phase, clustering phase, sentences are mapped to two-dimensional map using SOM map and then categorized hierarchally using GHSOM.
3. Finally, SOMvisua provides six visualization tools to represent clustering.

The main contribution of SOMvisua is the mixing of input graph model with the visual output graph model with **SOM** in an automatic way. SOMvisua proves that using graphs from the beginning to represent documents, produces a clustering process that is robust, and generates precise clustering.

SOMvisua provides tools to store visualizations in graphics file, export visualizations to encapsulated PostScript file, and export SOM to HTML file. User can adjust some global preferences for the visualization area such as background color, border size, and font size used for labels.

The SOMvisua server system is implemented in Java programming Language. It requires all similarity models and visualization to be loaded into memory. In the case of the one million sentences, 8 MB of memory is required to store all text features in memory.

SOMvisua User Interface: SOMvisua is a desktop application, which implements text clustering and visualization framework. The GUI is used as a text viewer. Through the GUI, the user could open test files, and add them to a list, which is used by the text viewer.

SOMvisua integrates four functionalities on top of the core SOMvisua that are done through SOMvisua GUI. These functionalities are:

- Converting text feature extraction method into similarity graph data structure.
- Visualizing the similarity graph with eight variant graph navigation algorithms.
- Performing SOM and GHSOM clustering for the similarity graph with visualization of SOM-grid.
- Six graphical visualization types of SOM and GHSOM clustering.

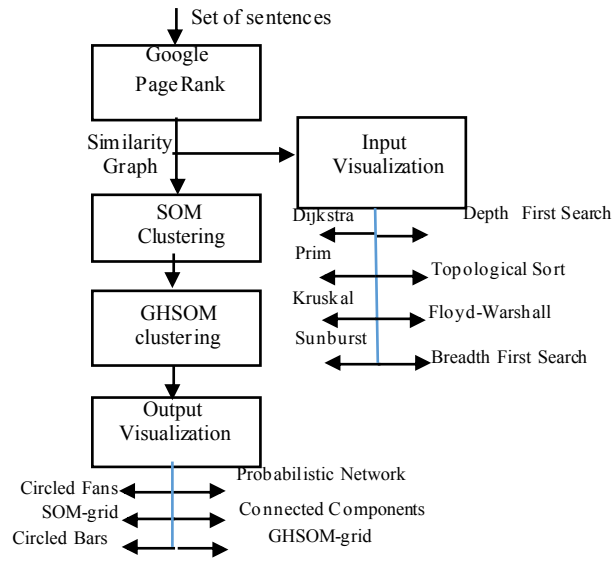

Figure 1: basic concept of SOMvisua

V. EXPERIMENTAL RESULTS

The section presents the experimental results of SOMvisua. We need to build the experimental framewok. Then, we implement the experimental framework, report query performance and show the graphical interface that allows easy visualizing of the selected text using SOMvisua.

V.1 Determining input text:

First step is to decide what input text data is used. To build the experimental framework, real text data is used. The data is fetched from an on-line Wikipedia web site, which offers free articles of texts for surfing purposes. Wikipedia has 4.3 million articles in English language, which considered meaningful text with feasible context and suitable as datasets. Each article is assigned to one or more topics, which enables us to perform a large-scale topic-evaluation experiment. In total, there are 11 different text topics in the website.

V.2 Selecting text similarity algorithm:

The next step in building the experimental framework is to select the text similarity algorithm to use for extracting features, to choose the parameters for representing data as graph method to achieve best clustering results, and to evaluate the configuration according to the measures we have used.

Due to the nature of human language, we used a content-based text similarity algorithm to demonstrate effectiveness of SOMvisua. We developed a mechanism to form graph data representation to be used as an input to SOM and GHSOM clustering algorithms. SOMvisua uses Google pageRank algorithm. Google pageRank algorithm is currently one of the text similarity algorithms with the best qualitative results.

The Google PageRank similarity function produces full representative similarity hierarchical graph. When PageRank is applied to all of the documents and the data is analyzed, the implications become much more interesting. For example, PageRank values are an exceedingly accurate map of user behavior probability. Additionally, SOM and GHSOM require the similarity Matrix technique to linearly combine its two words and concept similarity measures. However, with the Google pageRank method presented we have shown a way to use the complex similarity functions while still retrieving a very high percentage of concept representation.

V.3 Selecting clustering parameters in SOM:

To use the SOM and GHSOM clustering algorithms two parameters need to be determined: the map units per source and the map units per destination for the approximate graph sources and destinations in SOM. Both parameters have a direct impact on the clustering quality and the speed of the system. SOM configuration parameters are chosen via the dialog box shown in Figure 2.

V.4 Choosing the optimum set of parameters:

We conducted an experiment to evaluate how the approximate clustering method performs on the one million sentences using different parameter configurations. This experiment is the basis to choose the optimum set of parameters for the visualization system. To perform the experiment, we randomly selected 1000 sentences from the one million sentences, computed their exact 1–10 nearest neighbors in the whole article (of one million articles) and used SOM and GHSOM methods with different parameter settings to measure the impact on the clustering quality, comparing it to a manual exact clustering. In the experiment, the nearest neighbor is computed, measuring the percentage of true nearest neighbors compared to the exact solution.

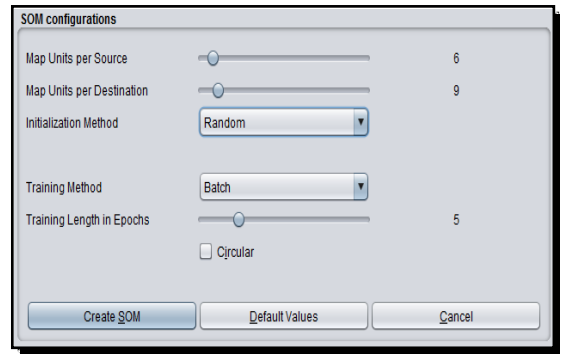

Figure 2: SOM configuration

The implementation of the actual clustering and visualization system is now straightforward. In an initial step, text document is analyzed and its similarity model is computed. All similarity models are allocated in memory. A graph data structure is built by Google pageRank. The number of nodes per graph depends on the actual number of words in the document. Google pageRank forms and stores ranking object to link between nodes. English language has a limited set of words and Google pageRank comprehensively evaluates words. After the document is analyzed, two

additional preprocessing steps are required before the system do visualization:

- Creating SOM cluster by selecting SOM parameters manualy. Then SOM enters a training phase to calculate epochs. Finally, SOM-Grid is displayed.
- Creating GHSOM cluster over the previously formed SOM.

GHSOM configurations dialog box will pop up as shown in Figure 3. In this step, more parameters are set. In addition to map units per source and map units per destination, growing and extension threshold should be adjusted. To measure the performance of SOMvisua, we run the system using a desktop PC with a core i7 CPU and 32GB of RAM memory.

Table 2 shows five different SOM and GHSOM configurations for one million sentences article. Figure 4 shows clustering execution time of these five SOM and GHSOM configurations. From the figure, we see that the final system is capable of visualizing text clustering in maximum of 0.29 seconds for a one million sentences article while returns about 94% of the correct nearest neighbors.

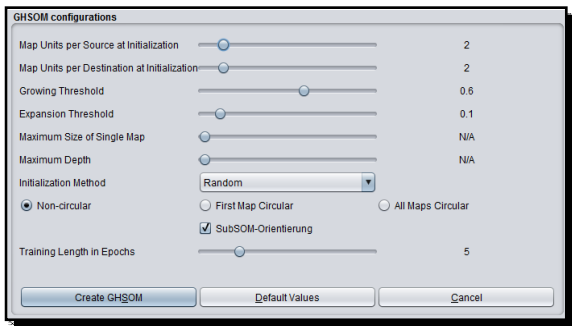

Figure 3: GHSOM configuration

The desktop application starts by displaying a menu, which is used to select a text document. Figure 5 shows a screenshot of the application. After selecting text document and the Google PageRank algorithm, SOMvisua starts to build similarity graph. Nodes construction will be displayed in the black messaging area. The text document that you choose will be displayed at right side text viewer area.

Similarity graph visualization

Similarity graph can be viewed in three forms: Graph, adjacency list, and adjacency matrix.

From "Extract Features" menu, you can choose "view input Graph" to start constructing eight tab panels. Figure 6 shows similarity graph visualized as nodes and vertices and ready to be tested.

SOMvisua provides a visual animation for algorithms execution with animation speed control. In the bottom of Figure 16->6 a slide bar controls the animation speed. The user can choose between directed graph or undirected graph. The results of the algorithm are viewed through the graph. Figure 6 shows Dijksrta's algorithm running to compute the shortest path to Node 5. Figure 7 shows kruskal's algorithm that finds a minimum spanning tree for a connected weighted graph. Figure 8 shows prim's algorithm that finds a minimum spanning tree for a connected weighted graph.

SOM and GHSOM clustering: To execute SOM clustering algorithm, the user selects "create SOM" submenu from SOM menu. SOM configurations dialog box is then displayed for the user to choose preferred configurations. "SOM-Grid" is then displayed in "SOM" tab as a sub tab of "output panel" tab as shown in Figure 9. "SOM" menu offers "save SOM" submenu to save SOM in a file to be loaded later using "load SOM" submenu. Another submenu "Export SOM to HTML" generates an HTML file that contains SOM-Grid. All previously mentioned options for SOM are also used for GHSOM clustering algorithm using "GHSOM" menu. GHSOM-Grid is shown in Figure 10.

Clustering Visualization: "visualization" menu contains five submenus for clustering visualization: basic circled bar, advanced circled bar, circled Fans, probabilistic network, and sun burst. These visualizations are shown in Figures 11, 12, 13, 14, and 15 respectively.

	Config 1	Config 2	Config 3	Config 4	Cofg 5
Map units per Source	2	4	6	8	10
Map units per destination	2	4	6	8	10
Growing threshold	0.2	0.4	0.6	0.8	1.0
Expanding threshold	0.2	0.4	0.6	0.8	1.0
Max size of single Map	20	40	60	80	100
Max depth	20	40	60	80	100
circularity	Non-circular	All maps	first	All maps	first

Table 2: five different SOM and GHSOM configurations

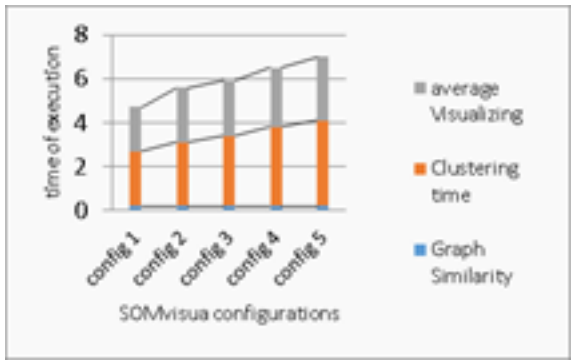

Figure 4: time to execute SOMvisua using five variant configurations

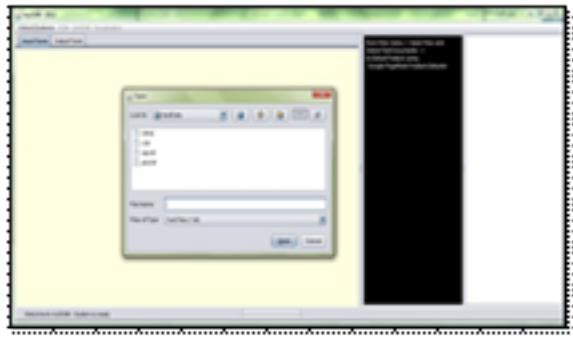

Figure 5: SOMvisua framework

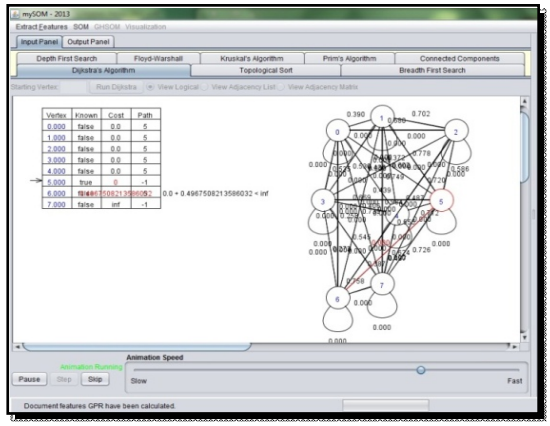

Figure 6:Dijksrta's algorithm

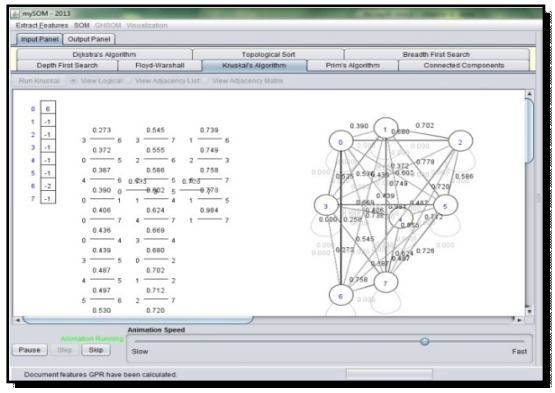

Figure 7:kruskal's algorithm

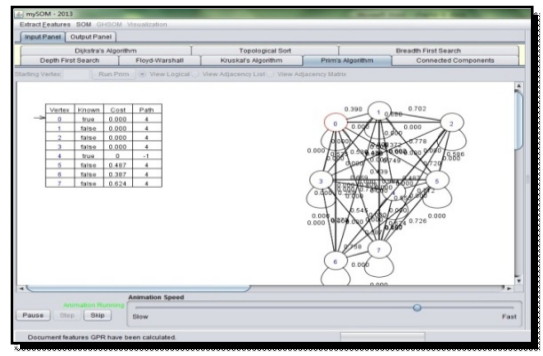

Figure 8:Prim's algorithm

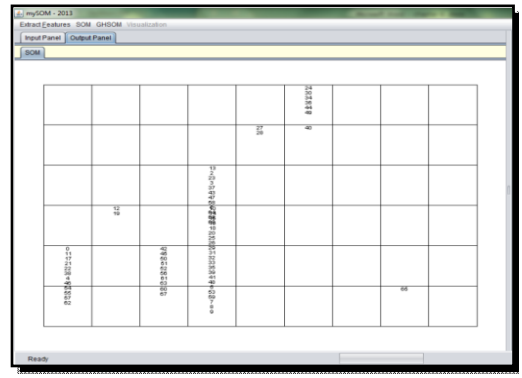

Figure 9: SOM-Grid preview

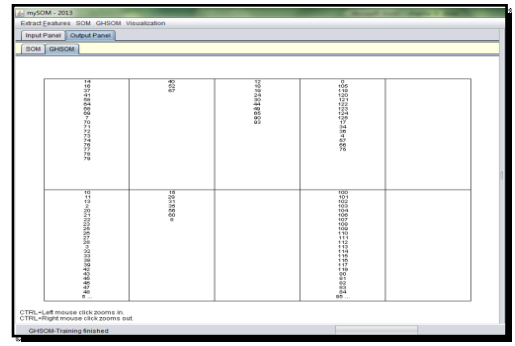

Figure 10:GHSOM-Grid preview

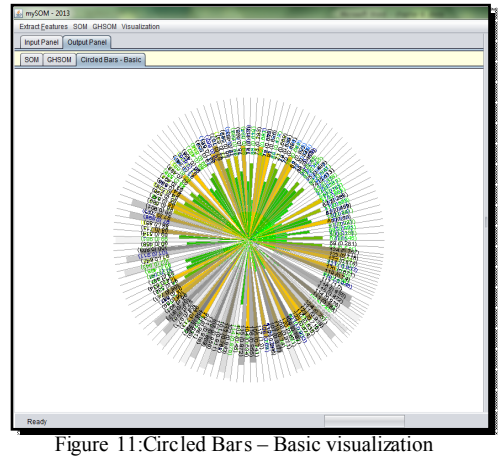

Figure 11:Circled Bars – Basic visualization

Figure 12:Circled Bars – advance visualization

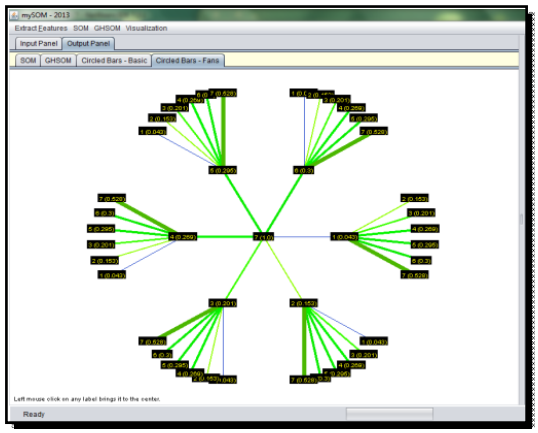

Figure 13: Circled Fans visualization

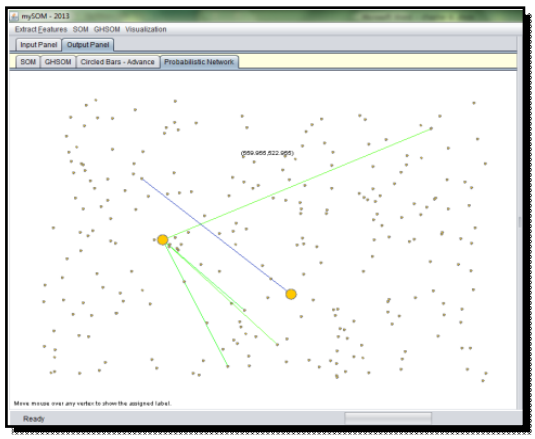

Figure 14: Probabilistic network visualization

I. CONCLUSION

We have designed and implemented a new framework called the SOMvisua for clustering sets of sentences of full-text article that is available in electronic form. SOMvisua is suitable for visualization of tasks in which the user has a vague

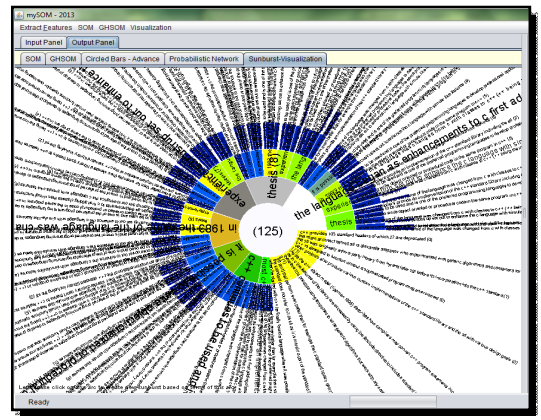

Figure 15: SunBurst visualization

idea of the contents of the text article being examined.With SOMvisua, the sentences are ordered meaningfully on a graph map according to their contents. Graph representation helps visualization by giving an overall view of what the information space looks like.

We showed how to use the text similarity features natively and correctly in SOM and GHSOM clustering algorithms, developed a method to alleviate the hub problem and created a visualizing solution for the reviewed class of text similarity algorithms. SOMvisua provides a visual animation for eight famous graph algorithms execution with animation speed control. SOMvisua provides a visual animation for eight famous graph algorithms execution with animation speed control. SOMvisua presents six types of visualization.

REFERENCES

[1] M. Schedl, P. Knees, and G. Widmer." Using CoMIRVA for Visualizing Similarities Between Music Artists". the 16th IEEE Visualization 2005 Conference , October 2005

[2] J. Aucouturier and F. Pachet. "Music Similarity Measures: What's the Use?". the 3rd International Symposium on Music Information Retrieval , pages 157-163 , October 2002.

[3] A. Langville and C. Meyer,"Google's PageRank and Beyond: The Science of Search Engine Rankings", Princeton University Press Princeton, ISBN:0691122024,July 2006

[4] http://en.wikipedia.org/wiki/PageRank

[5] Self Organizing Maps, Tom Germano March 23, 1999, http://davis.wpi.edu/~matt/courses/soms/, last accessed 24.July.2014

[6] F. Hussin, M. Farra and Y. Sonbaty, "Extending the Growing Hierarchal SOM for Clustering Documents in Graphs domain", International Joint Conference on Neural Networks, Pp. 4028–4035,June 2008.

[7] S. Kaski , T. Honkela , K. Lagus , and L. Kohonen , "WEBSOM self-organizing maps of document collections",Neurocomputing, vol. 21, May 1998

[8] The Growing Hierarchical Self-Organizing Map , Department of Software Technology, Vienna University of Technology

[9] The Growing Hierarchical Self-Organizing Map, Michael Dittenbach, Dieter Merkl, Andreas Rauber, Proceedings of the Int'l Joint Conference on Neural Networks (IJCNN'2000), Como, Italy, July 24-27, 2000, pp VI-15 - VI-19.

[10] D. Phuc and M. X. Hung, "Using SOM based Graph Clustering for Extracting Main Ideas from Documents", Research, Innovation and Vision for the Future International IEEE Conference Page(s) 209 - 214,July 2008

[11] A. Ultsch, "Self-organizing neural networks for visualization and classification", Information and Classification-Concepts Methods and Applications, Page(s) 307–313,September 1993.

[12] A. Shklovets and N. Axak, "Visualization of high-dimensional data using two-dimensional self-organizing piecewise-smooth Kohonen maps", Optical Memory and Neural Networks Journal, vol. 21, Page(s) 227-232 ,October 2012.

[13] R. Gruen and T. Kubota, , "A neural network approach to system performance analysis", IEEE Digital Object Identifier , Page(s) 349 - 354,April 2002.

[14] D. Merkl and A. Rauber, "Alternative ways for cluster visualization in self-organizing maps", 1st Workshop Self-Organizing Maps, Page(s) 106–111,May 1997.

[15] H. Yin, "ViSOM- a novel method for multivariate data projection and structure visualization", IEEE Transactions on Neural Networks, vol. 13,Page(s). 237–243, January 2002.

[16] E. Pampalk, A. Rauber, and D. Merkl, "Using smoothed data histograms for cluster visualization in self-organizing maps", International Work-Conference of Artificial Neural Networks, Page(s) 871–876, August 2002

[17] B. Choudhary and P. Bhattacharyya, "Text Clustering Using Universal Networking Language", Universal Networking Language Conference, vol. 16 , Page(s) 22-36, November 2001.

[18] H. Chim and X. Deng, "A New Suffix Tree Similarity Measure for Document Clustering", the 16th international conference on World Wide Web, Page(s) 121-130,May 2007

[19] A. Becks, S. Sklorz and M. Jarke, "A Modular Approach for Exploring the Semantic Structure of Technical Document Collection", the International Working Conference on Advanced Visual Interfaces, Page(s): 298-301,May 2000

[20] K. Lagus, T. Honkela, S. Kaski and T. Kohonen, "Self-organizing map of Document Collection, A New Approach to Interaktive Exploration", the 2nd International Conference on Knowledge Discovery and Data Mining, pages 238-243, August 1996

[21] X. Lin, D. Soergel and G. Marchionini, "A self-organizing Semantic Map for Information Retrieval", the 14th annual international ACM SIGIR conference on Research and development in information retrieval, Illinois, United States, pages 262-269,September 1991

[22] Mu-Chun Su1, Ta-Kang Liu and Hsiao-Te Chang, " improving the Self-Organizing Feature Map Algorithm Using an Efficient Initialization Scheme ", Tamkang Journal of Science and Engineering, Vol. 5, No. 1, pp. 35-48 (2002)

Mohammad A. Mikki is a professor of computer engineering at the Islamic university of Gaza with about twenty years of research, teaching, and consulting experience in various computer-engineering disciplines. Dr. Mikki got both his Ph.D. and Master of Science in Computer Engineering from Department of Electrical and Computer Engineering in Syracuse University in Syracuse, New York, USA in 1994 and 1989 respectively. He also got his Bachelor of Science in Electrical Engineering from the Department of Electrical Engineering at BirZeit, University in BirZeit in West Bank in 1984.

Khalid M. Kahloot is a computer Engineer at Department of Information Technology at Ministry of Education, North Directorate. Eng. Kahloot has 8 years experience in database administration and software developmet. Eng. Kahloot holds Master and Bachelor degrees in computer engineering from department of Electrical and Computer Engineering, Faculty of Engineering, Islamic University of Gaza sence 2014,2002 respectively.