Tracking English and Translated Arabic News using GHSOM

Ali Selamat[α] and Hanadi Hassen Ismail Mohammed[β]
[α]*Universiti Teknologi Malaysia, Malaysia*
[β]*Sudan University of Science and Technology, Sudan*

1. Introduction

After the September eleventh attacks, the media has contributed in clarifying the discrepancies between eastern and western cultures. As stated by Michael Binyon, a journalist in the The Times newspaper, whether they work for newspapers, television, or radio programs, all journalists are bound by their professional ethics codes, to provide truthful and accurate news. Unfortunately, only very few media outlets can boast of saying that everything they have put out is everything that had really happened (Michael Binyon, 2008). Take, for example, the violence in Gaza, a journalist was asked by the publisher of a Middle Eastern newspaper why the BBC gave so much coverage to the rockets aimed by Hamas at Israeli towns but gave so little coverage to the Israeli air strikes in Gaza. His answer was because there were no BBC correspondents based in Gaza. This shows that there is a need to track similar news from different resources because watching news from one resource may not always give audience the real picture of the event.

There are many advantages of finding similar content across different languages. For example, it can form the basis for multilingual summarization and the question answering support for web pages that provide questions and answers. It also facilitates comparative studies across national, ethnic, and cultural groups (Jin & Barrire 2005). Dittenbach, et al.,2001 have stated that human categorizations are based on grouping similar objects into a number of categories. This ahs been done in order to understand the differences between objects that belong to each defined category. There are many approaches have been used for finding similarity across multilingual documents such as neural networks, fuzzy clustering, genetic algorithms, support vector machines, etc. Most of these techniques are baed on the assumption on the availability of a clean corpus and the majority of these pages are written independently of each other (Jin & Barrire 2005). Therefore, two versions of the same topic that are written in two different languages cannot simply be taken as parallel corpora. One of the techniques used to solve this problem is Self-Organizing Map (SOM) (Kohonen, 1990). It is an unsupervised neural network algorithms which provide a topology-preserving mapping from a high-dimensional document space into a two dimensional map space. The documents that are on similar topics are located in neighboring regions (Dittenbach, et al.,2001). SOM produces additional information about the affinity or similarity between the clusters themselves by arranging them on a 2D rectangular or hexagonal grid when we cluster the documents. Similar clusters are neighbors in the grid, and dissimilar clusters are placed far apart in the grid (Zamir, 1999).

Similarity techniques have been used to find important information in the summarization of English news but it has not been used across languages (Chen et al., 2004). Finding similar news documents in Arabic and English news using machine learning algorithms is not an easy task. There are many news agencies nowadays that are broadcasting on what is happening around us and around the world such as Aljazeera (Aljazeera, 2007), Alsharq Alawsat (Alsharqalawsat, 2007), CNN (CNN, 2007), BBC (BBC, 2007), etc. Large networks provide more national and global news, while local stations concentrate more on the regional issues. Furthermore, even though everyone in the news industry claims their reporting is objective, the actual attitude in the broadcasting may be biased and different from network to network due to the background and cultural differences. Therefore, getting the same news from multiple sources provides the audience with more comprehensive views. These views would also be used by intelligence analyst in assessing counter-terrorism, political leadership, or country specific information.

In this paper, we analyze the appropriate techniques to find similar news across Arabic and English sources. This will provide the audience with multiple views of the broadcasted news because reading the news from a single source may not always reflect what is happening around the world due to different background, cultures, and opinions of the readers and writers. We have analyzed the similarity of the views on the news written in the news translations form Arabic and English texts using Self-organizing Map (SOM) (Kohonen, 1990). However, we have found that there are some difficulties in SOM that affect its performance. In order to improve its performance, we have used a Growing Hierarchical Self-Organizing Map (GHSOM) (Helmut, Dieter 2005). The research that has been undertaken is described in the rest of the sections.

2. Related Research On Arabic and English Translations

The discovering of new knowledge from textual information sources has increased the popularity in the information systems field. This is particularly important when an increasing availability of digital documents in various languages has enabled the web documents to be accessed by many internet users (Lee & Yang 2003).

To cross language boundaries between different languages, dictionaries are the most typical tools. However, the general-purpose dictionary is less sensitive in genre and domain and it is impractical to manually construct multilingual dictionaries for large applications. Corpus-based approaches, which do not have the limitation of dictionaries, provide a statistical translation model to cross the language boundary. Parallel corpora form the basis of much multilingual research in natural language processing (Lee & Yang 2003). There are different approaches for finding similar sentences across multiple languages. Most methods for finding similar sentences assume the availability of a clean parallel corpus. In some web pages that provide multiple languages, two versions of a page topic in two different languages are a good starting point for searching similar sentences. However, these pages may not always conform to the typical definitions of a bitext which current techniques assume.

Bitext generally refers to two versions of a text in two different languages (Adafre & Rijke, 2006). However, it is not known how the information is shared among the different languages in the same web page. Some pages tend to be the translations of each other. Thus, it is easy to use the automatic Parallel Text Identification (PTI) systems for aligning parallel web documents, which are direct translations of each other. The PTI systems crawl the Web to fetch potentially parallel multilingual Web documents (e.g Web spider) and to determine the parallelism between potential document pairs. Two modules are developed here. First, a filename comparison module is used to check the filename resemblance. Second, a content analysis module is used to measure the semantic similarity. The multilingual web documents retrieved by the web spider are initially passed to the filename comparison module to undergo a filename comparison process. Any two web pages in two different languages with their corresponding filenames resemble each other are picked up and aligned to form a parallel document pair. The multilingual web documents that remain unaligned after the filename comparison process will be passed to the content analysis module to carry out semantic content analysis (Chen et al., 2004).

3. Finding News Similarity Using SOM and GHSOM

3.1 News Collection

The available news documents were collected from several news websites such as Aljazeera news (Aljazeera, 2007), Alsharq Alawsat (Alsharqalawsat, 2007), etc. We have used the news crawler software (Selamat A.et al., 2007), available from the internet in order to get the desired URLs and then download the news materials. The Arabic news collection consists of the translated version of the news.

3.2 Documents Preprocessing

Documents preprocessing refers to the act of cleaning the text and processing the data before it is parsed. It consists of three main operations. Firstly, the conversion of HTML files to plain text files, where the downloaded HTML files were striped from all HTML-tags, and then converted to plain text files using the available public HTML-to-text converters. Secondly, the removal of stop words from the documents, where all the words which are not important to retrieval will be removed because they exist in any document repeatedly and do not affect the meaning. In order to remove the stop words, a file containing a large number of common stop words is used.

The news text words are matched with the words in the stop words list. If this matches, the word is automatically excluded from the text. Thirdly, the stemming operation, where prefixes and suffixes will be removed to obtain word stems. This can be achieved using a stemming program. The stemming is highly language-dependent. For the English language, the well-known and well-established Porter's stemmer (Selamat A. & Omatu S, 2004) provides high quality stemming for content representation purposes. The translated Arabic news contents will also be stemmed using the same stemmer used for the English content.

3.3 Documents Parsing

Parsing is the process of creating a vector representation of the texts that can be used for training a self-organizing map. A list of all words occurring in the document collection is called the template vector. To obtain the individual vectors, every document is described by the words occurring in it. In the simplest form of document representation, a binary indication is used for describing the fact whether a word is present or not, leading to a corresponding vector representation of 0 or 1. This type of representation is usually referred to as binary representation. The importance of every word in each document is based on the term frequencies that will be obtained using the so-called term frequency time's inverse document frequency (tf x idf) representation, where the importance of each word is judged with respect to the number of documents it appears in, as appointed in chapter two.

3.4 Creating feature vectors

A numerical representation of the documents collection was created in order to be able to automatically organize documents by content. This representation is created by full text indexing the documents. The template vector is the list of all words occurring in the document collection. This template vector consists of a very huge number of words. For better representation, this list should be pruned to a subset relevant and this can be achieved by removing all words that appear in a very large number (e.g., more than 90%) of documents, as these words do not contribute to content separation. This process also includes removing the stop words such as *a*, *the*, *you*, *we*, and *they*. The advantages of this process are the non-significant words are removed and the total size of the document file is reduced. Words that appear in only very few documents were also be removed.

3.5 Growing Hierarchical Self-organizing Maps

The Growing Hierarchical Self-Organizing Map (GHSOM), which is an extension to the growing grid SOM and hierarchical SOM, has been used to build a hierarchy of multiple layers where each layer consists of several independent growing SOMs. The size of these SOMs and the depth of the hierarchy are determined during its learning process according to the requirements of the input data. For the initial setup of the GHSOM, at Layer 0, a single-neuron SOM is created and the neuron's weight vector is initialized as the average of all the input vectors. Then, the learning process starts at Layer 1 with a small SOM (usually a grid) whose weight vectors are initialized to random values. The GHSOM grows in two dimensions: horizontally (by increasing the size of each SOM) and hierarchically (by increasing the number of layers).

For the horizontal growth, shown in Fig. 1, each SOM modifies itself in a systematic way, very similar to the growing grid so that each neuron does not represent too large of an input space. For the hierarchical growth, shown in Fig. 2, the principle is to periodically check whether the lowest layer SOMs have achieved sufficient coverage for the underlying input data (Tangsripairoj, Samadzadeh 2005). The basic steps of the horizontal growth and the hierarchical growth of the GHSOM are summarized according to Tangsripairoj et al.(Tangsripairoj, Samadzadeh 2005).

The growth process of the GHSOM is controlled by the following four important factors. First, the quantization error of a neuron i, qe_i is calculated as the sum of the distance between the weight vector of neuron i and the input vectors mapped onto this neuron.

Second, the mean quantization error of the map (mqe_m) is the mean of all neurons quantization errors in the map. Third, the threshold τ_1 is for specifying the desired level of detail that is to be shown in a particular SOM. Fourth, the threshold τ_2 is for specifying the desired quality of input data representation at the end of the learning process.

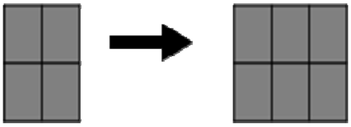

Fig. 1. GHSOM addition of new column

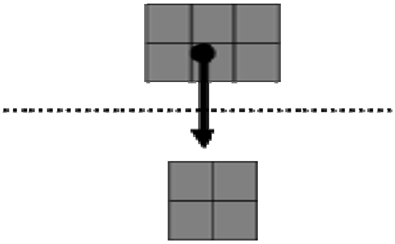

Fig. 2. GHSOM expansion in a sub layer

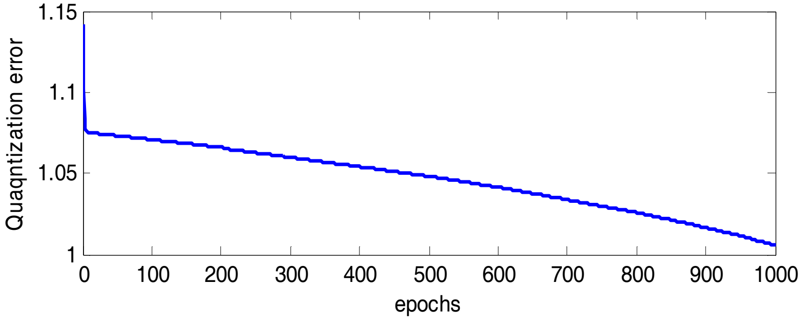

Fig. 3. Quantization Error for each epch

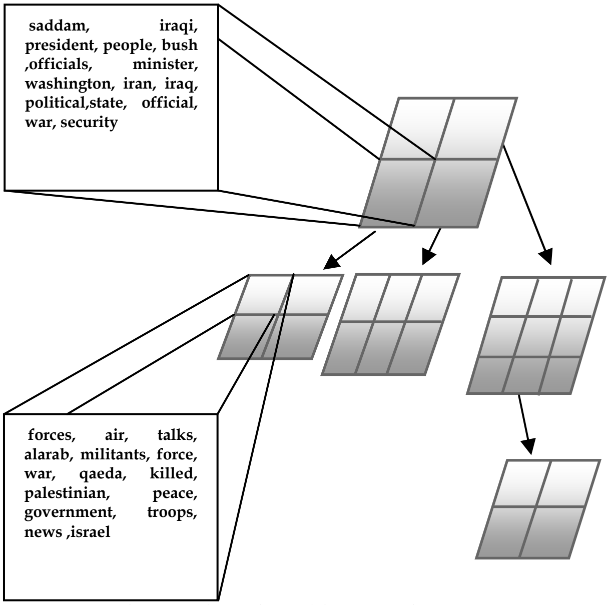
saddam, iraqi, president, people, bush ,officials, minister, washington, iran, iraq, political,state, official, war, security
forces, air, talks, alarab, militants, force, war, qaeda, killed, palestinian, peace, government, troops, news ,israel

Fig. 4. 3-Layer GHSOM map with sample political keywords

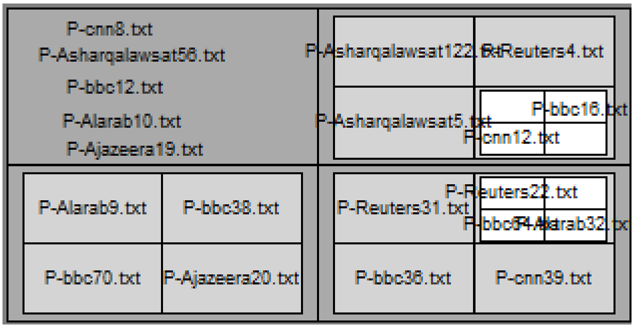

Fig. 5. 3-Layer GHSOM map with political documents cluster

4. Discussion of Results

4.1 Map quality

For exploratory data analysis tasks where the SOM and GHSOM are used as tools to display similarity relationships from a high-dimensional input space on a low dimensional mapping space, the quality of this mapping is essential. The quality of data representation is measured in terms of the deviation between a units model vector Best Matching Unit (BMU) and the subset of input data vectors represented by this unit. The deviation can be calculated as either the mean quantization error (*mqe*) or the quantization error (*qe*) of the single unit. This error is used to measure the map resolution as mentioned by Liu et al. (Liu, et al., 2005).

$$\left[mqe_i = \frac{1}{n} \bullet \sum_j \left\| m_i - x_j \right\| \right] \tag{1}$$

Where m_i is the codebook vector for *unit i* , x_j for all *j* are the input vectors mapped on *unit i*, and n is the number of *xi*. Referring to Fig. 5, using different vector size, the GHSOM achieves better results or less quantization error in several vector sizes than SOM. This means that GHSOM has the ability to make better map representation than SOM.

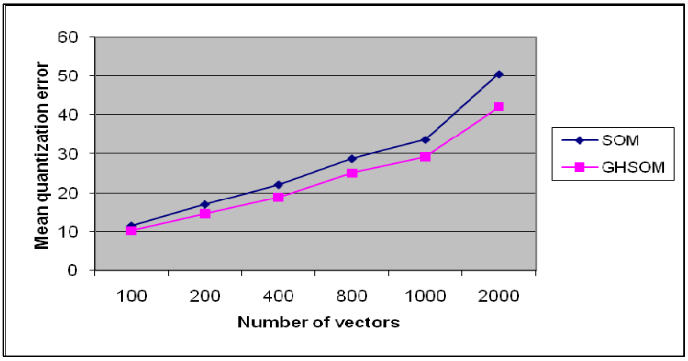

Fig. 6. Mean Quantization Error for SOM and GHSOM.

4.2 Documents Similarities Measurements

In the context of document clustering, the ability of an algorithm to classify a document into one or several categories is of high interest to the user. The classification effectiveness of this study has been measured using standard information retrieval measurements that are Precision(P), Recall(R), and F1. These can be expressed as:

$$\Pr ecision = \frac{a}{a+b} \tag{2}$$

$$\operatorname{Re} call = \frac{a}{a+c} \tag{3}$$

$$F1 = \frac{2PR}{P+R} \tag{4}$$

Category set C = $c_1, c_2, \ldots, cn$		Expert Judgement	
		Yes	No
System Judgement	Yes No	a c	b d

Table 1. Exploration of Measurements Used In the Experiments

Four main categories have been used to train the networks that are business, politics, sports, and technology. The Precision, Recall, and F1 measures of using SOM and GHSOM that have been calculated for all of these categories are shown in Table2, Fig. 6,7, and 8. Plots for these categories in Fig. 6 and 7 show that the GHSOM algorithm has consistently achieved higher Precision and Recalls levels than the standard SOM. Referring to Fig. 6 and 7, we have found that the average of both the Precision and Recall measures for SOM and GHSOM are 87% and 93%, respectively.

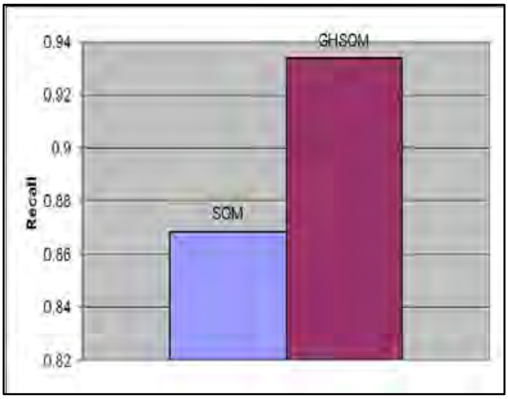

Fig. 7. Average Precision for SOM and GHSOM.

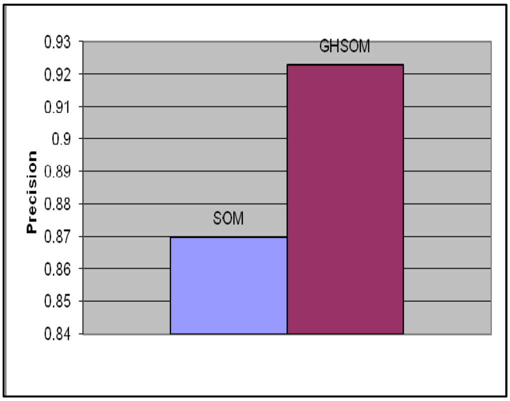

Fig. 8. Average Recall for SOM and GHSOM.

Figure 8 shows the achievable F1 measures for all categories. It is defined as the harmonic mean of Precision and Recall as shown in Eqs. (2), (3), and (4), respectively, and the yields values in the interval [0, 1], with F1 = 0 when no relevant documents are found, and F1 = 1 when all documents from a given class are retrieved with no errors. The detail explanation on each category of document similarity is shown in Table2. As shown in the table, the GHSOM outperforms the original SOM algorithm in all clusters except at 50 documents when the precision is the same in Business and Sports clusters. Referring to Fig. 8, we have found that the average of F1 measures for SOM and GHSOM are 87% and 92%, respectively. It shows that the GHSOM algorithm has performed better in all categories.

Precision	Business		Politics		Sports		Technology	
	SOM	GHSOM	SOM	GHSOM	SOM	GHSOM	SOM	GHSOM
At 50 docs	1	1	0.88	1	1	1	0.75	1
At 100 Docs	0.96	1	0.88	1	0.88	1	0.82	1
At 500 Docs	0.94	0.96	0.89	0.98	0.95	0.97	0.59	0.83
At 1000 Docs	0.90	0.96	0.89	0.94	0.98	0.93	0.70	0.87

Table 2. Precision at 1000 documents of four clusters.

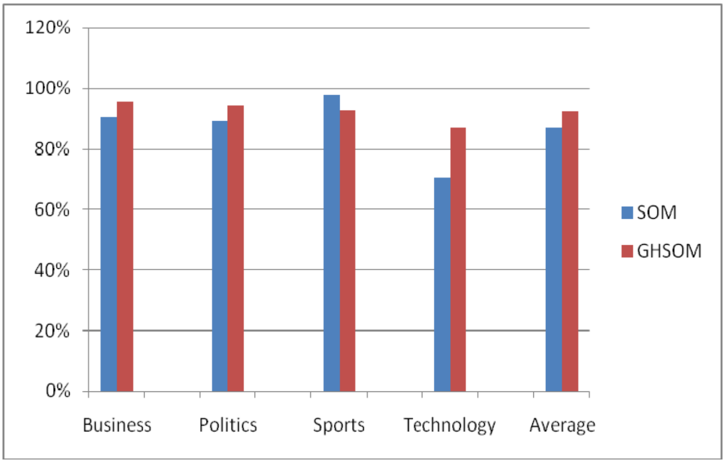

Fig. 9. F1 measures for four categories achieved by SOM and GHSOM.

Average of F1 measures for SOM and GHSOM are 87% and 92%, respectively.

4.3 The resulting map

The SOM and GHSOM maps consist of cells where each cell contains similar content documents, as shown in Fig. 4. The keywords appear in the figure indicate that these key keywords have higher frequency. Thus, most of the documents containing these keywords will be mapped in these cells, as shown in Fig. 5. In order to compare the results of the SOM and GHSOM applications, both algorithms were trained using large set of documents (1000 documents). Because of the limited space, we do not show the full maps that contain all the documents. As shown in Fig. 10, the SOM map contains the four main categories that are business, politics, sports, and technology as labeled with cluster titles. News with similar features are apparently located on the nearby regions of the map. The problem with the SOM map is that the map size should be predetermined and whenever the input space becomes larger, it is getting difficult to clearly visualize the map and the clusters areas. In the GHSOM resulted map illustrated in Fig. 11, it can be seen that the clusters are the areas with high densities on the map that were further hierarchically expanded by the growing

SOMs. In the figure, the top layer maps are depicted in gray and the bottom layer maps are depicted in white.

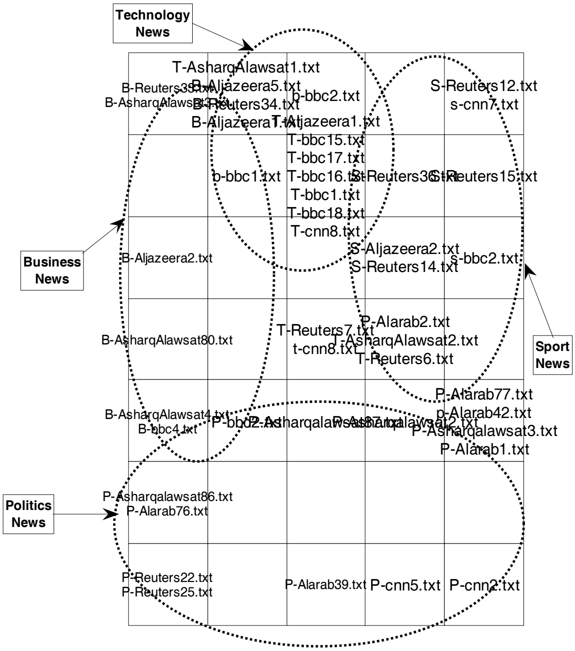

Fig. 10. The resulting 7X5 SOM map.

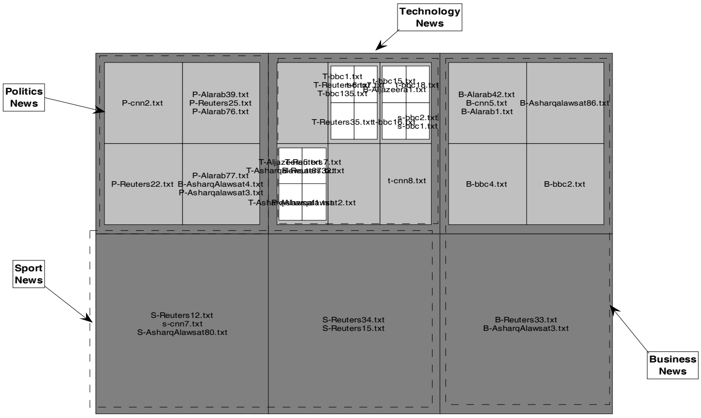

Fig. 11. The resulting 3-Layer GHSOM map.

5. Conclusion

Based on the experiments performed in this research, it can be concluded that the quantization error is a map quality measure which is data-dependent. It measures the map in terms of the given data. Typically, the quality of the map is measured in terms of the training data. From the experiments, lesser quantization error was obtained by GHSOM than by SOM which means that the data sent to the GHSOM is well located in the map. When training SOM and GHSOM, the training time taken by using the SOM algorithm was longer compared to the training time achieved by the GHSOM, as the factor of vector size increases. This indicates that the SOM algorithm is not useful for large number of training set documents. As the SOM algorithm cannot handle the hierarchies of the data, it is much difficult for the user to keep an overview of the various clusters. Also, the data hierarchies could not be discovered by the algorithm. We have evaluated the classification quality of SOM and GHSOM using Precision, Recall and F1 measures based on the results obtained by the algorithms. The results have shown that the GHSOM performed better than the SOM in Precision and Recall results for the four main clusters that were used in the training data. Also, when calculating the F1 measures, the GHSOM obtained better results than the SOM. The inclusion of word sense disambiguation on Arabic script documents that will be used to identify the semantic of documents is subject for future research.

6. References

Aalarab (2007). Aalarab news. www.alarab.com. 2007.

Adafre, Rijke (2006). Finding Similar Sentences across Multiple Languages in Wikipedia. 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 62-69, Trento, Italy, April 2006, Association for Computational Linguistics, Morristown, NJ, USA.

Aljazeera (2007). Aljazeera news. www.aljazeera.net. 2007.

Asharqalawsat (2007). Asharqalawsat news. www.asharqalawsat. 2007.

BBC (2007). BBC news. www.bbc.com. 2007.

Chen, R. Chau, C. Yeh (2004). Discovering Parallel Text from the World Wide Web. Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation, pp 157–161, Dunedin, New Zealand, 2004, Australian Computer Society, Inc. Darlinghurst, Australia.

CNN (2007). CNN news. www.cnn.com. 2007.

Dittenbach, Rauber, D. Merkl(2001). Business, Culture, Politics, and Sports - How to Find Your Way Through a Bulk of News? On Content-Based Hierarchical Structuring and Organization of Large Document Archives. Proceedings of the 12th P International Conference on Database and Expert Systems Applications, pp 200 – 210, 3-540-42527-6, Munich, Germany, September 2001,Springer-Verlag London, UK.

Evans (2005). Identifying Similarity in Text: Multi-Lingual Analysis for Summarization. Ph.D. Thesis. Columbia University. 2005.

Helmut, Dieter (2005). A comparison of support vector machines and self-organizing maps for e-mail categorization. Proceedings of the 4th Australasian Data Mining Conference (AusDM'05), pp 189-204, 1-86365-716-9, Sydney, Australia, December 2005, University of Technology Sydney, Australia.

Jin, Barrire (2005). Exploring sentence variations with bilingual corpora. Corpus Linguistics 2005 conference. Birmingham, United Kingdom, July 2005, NRC Institute for Information Technology, Canada.

Kohonen (1990). The self-organizing map. Proceedings of the IEEE, pp 1464 – 1480, 0018-9219, September 1990, IEEE.

Lee, Yang (2003). A Multilingual Text Mining Approach Based on Self-Organizing Maps. Journal of applied intelligent. Volume 18, No.3, (May 2003), page numbers (295-310), 0924-669X.

Liu, Wang, Zheng (2005). Mental tasks classification and their EEG structures analysis by using the growing hierarchical self-organizing map. Neural Interface and Control, 2005, Proceedings of the First International Conference, 115- 118, 0-7803-8902-6, May 2005, IEEE.

Michael Binyon(2008). *September 11th and the Western Media and Cross - Cultural Misunderstanding Role of Dialogue between Arab And West Seminar*, Kuwait, 2008.

Selamat A. and Omatu S. (2004). Feature Selection and Categorization of Web Pages Using Neural Networks, Int. Journal of Information Sciences, Elsevier Science Inc. Vol. 158, (January 2004), page number (69-88).

Selamat A., Choon N-C, Abu Bakar A.Z., Mikami Y.(2007), Arabic Script Web Documents Language Identification Using Decision Tree-ARTMAP Model, 2007 International Conference on Convergence Information Technology (ICCIT 2007), pp. 21-23, November 2007, Gyeongju-si, Gyeongbuk, Korea.

Tangsripairoj, Samadzadeh (2005). Organizing and visualizing software Repositories using the growing hierarchical Self-organizing map. Proceedings of the 2005 ACM symposium on Applied computing SAC, pp. 1539- 545, 1-58113-964-0, Santa Fe, New Mexico, 2005, ACM, New York, NY, USA.

Xafopoulos A., Kotropoulos C., Almpanidis G. and Pitas I(2004). Language Identification in Web Documents Using Discrete HMMs. Pattern Recognition. Vol. 137, No. 3, March2004, page numbers(583-394), 0031-3203.

Zamir (1999). Clustering Web Documents:A Phrase-Based Method for Grouping Search Engine Results. University of Washington: Ph.D.Thesis.

Zhai, Shah (2005).Tracking News Stories Across Different Sources.Proceedings of the 13th annual ACM international conference on Multimedia, pp. 2-10, 1-59593-044-2, Hilton Singapore, 2005, ACM, New York, NY, USA.