J. Intelligent Learning Systems & Applications, **2010, 2, 117-125**
doi:10.4236/jilsa.2010.23015 Published Online August 2010 (http://www.SciRP.org/journal/jilsa)

Scientific Research

A SOM-Based Document Clustering Using Frequent Max Substrings for Non-Segmented Texts

Todsanai Chumwatana, Kok Wai Wong, Hong Xie

School of Information Technology, Murdoch University, South St, Murdoch, Australia.
Email: {T.Chumwatana, K.Wong, H.Xie}@Murdoch.edu.au

Received March 25th, 2010; revised July 15th, 2010; accepted July 30th, 2010.

ABSTRACT

This paper proposes a non-segmented document clustering method using self-organizing map (*SOM*) *and frequent max substring technique to improve the efficiency of information retrieval. SOM has been widely used for document clustering and is successful in many applications. However, when applying to non-segmented document, the challenge is to identify any interesting pattern efficiently. There are two main phases in the propose method: preprocessing phase and clustering phase. In the preprocessing phase, the frequent max substring technique is first applied to discover the patterns of interest called Frequent Max substrings that are long and frequent substrings, rather than individual words from the non-segmented texts. These discovered patterns are then used as indexing terms. The indexing terms together with their number of occurrences form a document vector. In the clustering phase, SOM is used to generate the document cluster map by using the feature vector of Frequent Max substrings. To demonstrate the proposed technique, experimental studies and comparison results on clustering the Thai text documents, which consist of non-segmented texts, are presented in this paper. The results show that the proposed technique can be used for Thai texts. The document cluster map generated with the method can be used to find the relevant documents more efficiently.*

Keywords**:** *Frequent Max Substring, Self-Organizing Map, Document Clustering*

1. Introduction

Document clustering has been an important issue [1] due to the rapid growth in the number of electronic documents. Document clustering, sometimes can be generalized as text clustering, indentifies the similarity of documents and summarize a large number of documents using key attributes of the clusters. Document clustering uses unsupervised learning techniques and may assist fast information retrieval or filtering [2]. This is because clustering technique categorizes documents into groups based on their similarity in term of their member occurrences. Thus clustering can be used to categorize document databases and digital libraries, as well as providing useful summary information of the categories for browsing purposes. In information retrieval, a typical search on document database or the World Wide Web can return several thousands of documents in response to the user's queries. It is often very difficult for users to identify their documents of interest from such a huge number of documents. Clustering the documents enables the user to have a clear and easy grasp of the relevant documents from the collection of documents which are similar to each other and could be relevant to the user's queries.

For text clustering in information retrieval, a document is normally considered as a bag of words, even though a document actually consists of a sequence of sentences and each sentence is composed from a sequence of words. Very often the positions of words are ignored when performing document clustering. Words, also known as indexing terms, and their weights in documents are usually used as important parameters to compute the similarity of documents [3]. Those documents that contain similar indexing terms and frequencies will be grouped under the same cluster. This process is straightforward for European languages where words are clearly defined by word delimiter such as space or other special symbols. European texts are explicitly segmented into word tokens that are used as indexing terms. Many algorithms have been developed to calculate the similarity of documents and to build clusters for fast information retrieval [4]. In contrast, document clustering can be a challenging task for

many Asian languages such as Chinese, Japanese, Korean and Thai, because these languages are non-segmented languages, *i.e.*, a sentence is written continuously as a sequence of characters without explicit word boundary delimiters. Due to this characteristic, texts in a non-segmented document cannot be directly used to calculate the similarity. Some preprocessing needs to be performed first to discover keywords for Asian documents before clustering. As a result, most approaches for clustering non-segmented documents consist of two phases: a text mining process to extract the keywords, and a document clustering process to compute the similarity between the input documents.

2. Keyword Extraction

Keywords are usually regarded as an important key to identifying the main content of the documents. Most of the semantics are usually carried by nouns, although a sentence in a natural language text is composed of nouns, pronouns, articles, verbs, adjectives, adverbs, and connectives. Keyword extraction is one of the main applications of text mining. The objective of text mining is to exploit useful information or knowledge contained in textual documents [5]. Information Extraction (IE) is an essential task in text mining that describes a process of discovering interesting keywords underlying unstructured natural-language texts. Most keyword extraction methods proposed in the literature were accomplished by constructing a set of words from given texts. Keywords will then be selected from the set of words during the preprocessing step. Many approaches have been proposed to extract keywords from non-segmented documents such as Chinese [6], Japanese [7] or Thai documents [8]. Most techniques are based on word segmentation which is one of the most widely used information extraction techniques in Natural language Processing (NLP). However, most word segmentation approaches involve complex language analysis and require long computational time. After keyword extraction is performed, keywords are then transformed into feature vector of the words that appear in the documents. The term-weights (usually term-frequencies) of the words are also contained in each feature vector. The vector space model (VSM) has been a standard model of representing documents by containing the set of words with their frequencies [1]. In the VSM, each document is replaced by the vector of the words. The vector size is dependent on the number of keywords that appear in the documents. For instance, let w_{ik} be the weight of keyword k that appear in the document i, and $D_i = (w_{i1}, w_{i2},\ldots, w_{it})$ is the feature vector for document i, where t is the number of unique words of all documents. Therefore, the size of the feature vector is equal to t dimension as shown in **Figure 1**.

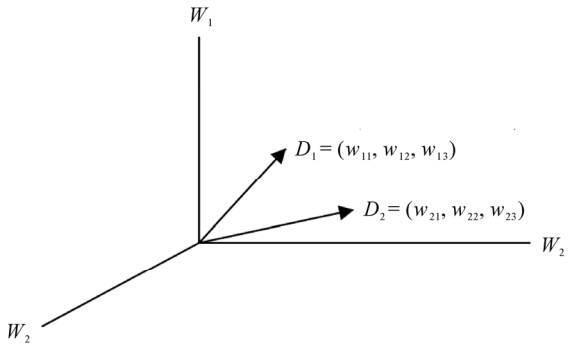

Figure 1. The example of the document vectors in 3-dimension

From **Figure 1**, the similarity between two documents can be computed with one of several similarity measures based on two corresponding feature vectors, e.g., cosine measure, Jaccard measure, and Euclidean distance measure [9].

3. Document Clustering Algorithms

In document clustering, there are two main approaches: hierarchical and partitional approaches [10,11,4]. The hierarchical approach produces document clusters by using a nested sequence of partitions that can be represented in the form of a tree structure called a dendrogram. The root of the tree contains one cluster covering all data points, and singleton cluster of individual data point are shown on the leaves of the tree. There are two basic methods when performing hierarchical clustering: agglomerative (bottom up) and divisive (top down) clustering [4]. The advantages of hierarchical approach are that it can take any form of similarity function, and also the hierarchy of clusters allows users to discover clusters at any level of detail. However, this technique may suffer from the chain effect, and its space requirement is at least quadratic or $O(n^2)$ compared to the k-means algorithm that provide $O(Iknm)$ where I is the number of necessary iterations, k is the number of clusters, n is the number of documents and m is the dimensionality of the vectors. The partitional approach [12], on the other hand, can be divided into several techniques, e.g., k-means [13], Fuzzy c-means [14], QT (quality threshold) [15] algorithms., The k-means algorithm is more widely used among all clustering algorithms because of its efficiency and simplicity. The basic idea of k-means algorithm is that it separates a given data into k clusters where each cluster has the center point, also called centroid, which can be used to represent the cluster. The main advantages of k-means algorithm are its efficiency and simplicity. Its weaknesses are that it is only applicable to data sets where the notion of the mean is defined, the number of clusters can be identified by users, and it is sensitive to data points that are very far away from other points called outliers [1]. Fur-

thermore, Self-organizing map (SOM) [16,17] can be used as one of the clustering algorithms in the family of an artificial neural network. The self-organizing map is unsupervised neural network, capable of ordering high dimensional data in such a way that similar inputs are grouped spatially close to each other. To use SOM in document clustering, text documents are described by features with high dimensionality, and SOM based techniques have been successfully applied to document clustering. Some of the successful applications of SOM in document clustering are described in the next section.

4. Related Works

Many clustering techniques have been developed and can be applied to clustering English documents Most of these traditional approaches use documents as the basis for clustering [18,19]. The Vector Space Document (VSD) model is a very widely used data representation model for document clustering [20]. This data model starts with a representation of any document as a feature vector of the words that appear in documents. The term-weights of the words are also contained in the feature vectors. The similarity measures are used to compute the similarity of two document vectors. An alternative approach of document clustering is phrase-based document clustering. Zamir and Etzioni [21] introduced the notion of phrase-based document clustering. They proposed to use a generalized suffix-tree to obtain information about the phrases between two documents and use common phrases to cluster the documents. According to [22], Bakus, Hussin, and Kamel used a hierarchical phrase grammar extraction procedure to identify phrases from documents and used these phrases as features for document clustering. The self-organizing map (SOM) method was used as the clustering algorithm. An improvement in clustering performance was demonstrated when using phrases rather than single words as features.

Mladenic and Grobelnik used a Naive Bayesian method to classify documents based on word sequences of different length [23]. Experimental results show that using the word sequences whose length is no more than 3 words can improve the performance of a text classification system. But when the average length of used word sequences is longer than 3 words, there will be no difference between using word sequences or single words.

However, there are not many research works on phrase-based document clustering for Asian languages, primarily due to the fact that most Asian language texts are non-segmented and it is difficult to separate words and phrases from the non-segmented texts. Most document clustering approaches require a preprocessing stage where word segmentation, stopword removal or semantic analysis are performed. NLP techniques provide good support for this step. Word segmentation is very important step involved in most NLP processing tasks. A text is separated into a sequence of tokens by using word segmentation techniques. Many approaches have been proposed for Asian languages such as Chinese, Japanese, Korea and Thai languages.

In [24], a Chinese document clustering method using data mining technique and neural network model was proposed. This technique was divided into two main parts: the preprocessing part which provides Chinese sentence segmentation method, and the clustering part that adopts the dynamical SOM model with a view to dynamically clustering data. In addition, this method uses term vectors clustering process instead of document vectors clustering process.

In Thai language, Canasai and Chuleerat propose a parallel algorithm for clustering text documents based on spherical *k*-means [25]. They implemented an algorithm on the PIRUN Linux Cluster, which is a parallel computer using cluster computing technology. Experimental results show that the use of parallel algorithm can significantly improve clustering performance.

5. A SOM Based Clustering Using Frequent Max Substrings for Non-Segmented Texts

In this section, we describe a new method that combines Kohonen's SOM and frequent max substring technique to process the non-segmented text documents into clusters. SOM is one of the main unsupervised learning methods in the family of artificial neural networks (ANN) that was first developed by Teuvo Kohonen in 1984 [26]. The SOM can be visualized as a regular two-dimensional array of cells or nodes (neurons). The SOM algorithm defines a mapping from the input vector onto a two-dimensional array of nodes. When the input vector $x(t) \in R^n$ is given, it is connected to all neurons in the SOM array denoted as vector $m_i(t) \in R^n$,which are associated by each neuron and is gradually modified in the learning process. The input vector $x(t) \in R^n$ is the input data sets where t is the indexing terms of the input documents. These input data sets have to be mapped with all neurons in the map that is denoted as two-dimensional network of cells or the model vector $m_i(t) \in R^n$.

In mapping, the node where vector m_i is most similar to the input vector x will be activated. This node is often called a best-matching node or a winner. The winner and a number of its neighboring nodes in the SOM array are then turned towards the input vector x according to the learning principle.

The frequent max substring technique is an information extraction technique used to identify the terms called frequent max substrings from non-segmented texts where the word boundary and characteristic are not clearly defined. This technique was first introduced in 2008 [27] and has been proposed as an alternative language-independent technique for keyword extraction for non-segmented texts [28,29]. It also has been applied in application for

indexing for non-segmented languages [8,30] as this technique provides good significant in term of the efficiency of storage space which could be able to support the rapid growth in the number of electronic non-segmented documents. The frequent max substring technique classifies indexing terms, known as frequent max substrings, from the non-segmented texts where the word boundaries are not clearly defined. The frequent max substrings refer to the substrings that appear frequently (at a given frequency threshold value *f*) and have the maximum length on the given string, so these terms are likely to be the patterns of interest. We extract the set of frequent max substrings or *FMAX* by using the frequent max substring technique. This technique uses Frequent Suffix Trie or FST data structure to explore the indexing terms [27]. The FST data structure is similar to suffix trie structure, that is an efficient substring enumeration method. However, FST data structure enumerates substrings with their frequencies and positions information while suffix trie structure enumerates only substrings without their frequencies information. Therefore, we employ FST data structure in order to support extracting the frequent max substrings. In this technique, the parameter and the predetermined frequency are applied to reduce the number of the indexing terms. This method uses the two reduction rules: 1) reduction rule using the predetermined frequency to check extracting termination, 2) reduction rule using superstring definition to reduce the number of indexing terms extracted. This technique also uses heap data structure to support computation [27].

In this paper, we use a set of non-segmented documents (Thai documents) as an input to train a map using SOM. We will describe the process of clustering as follows.

Let *D* be a document collection consisting of *n* documents, d_1, d_2, ..., d_n. Firstly, we use the frequent max substring technique [27] to generate a set of frequent max substrings at the given frequency threshold value *f* from the document collection to be used as the set of indexing terms for the document collection.

Assuming the above process produces *m* frequent max substrings from the document collection, denoted *FMAX* = (fm_1, fm_2, ..., fm_m). where fm_i is the *i*th frequent max substring generated from the document collection. These *m* substrings are used as our indexing terms.

We will then calculate the weight w_{ij} which represents the frequency of indexing term fm_i occurring in document d_j for each indexing term and each document. Finally we construct an $m \times n$ matrix of such weights. In this matrix, row *i* represents the frequencies of occurrence of the *i*th indexing term fm_i in the *n* documents, while *j*th column represents the document vector for document *j*, as depicted in **Figure 2**.

Figure 2 shows an example of document matrix. In a document matrix, each element w_{ij} is at least at *f* if fm_i occurs in the document d_j or 0 if fm_i does not appear in the document d_j, *i.e.*,

$$w_{ij} = \begin{cases} \geq f & \text{if } fm_i \text{ occurs in } d_j \\ 0 & \text{otherwise} \end{cases}$$

After the document matrix is obtained, the document vectors are presented to SOM for clustering. These documents can be labeled into neurons according to the similarity of their document vectors. Two documents containing the same or similar document vectors will map to the same neuron. In contrast, the documents may map to distant neurons if they contain different or non-overlapping frequent max substrings. Furthermore, the documents with similar frequent max substrings may map to neighbouring neurons. This means that the neurons can form the document clusters by examining mapped neurons in the document cluster map. We depict the organization of the document map that clusters similar documents into the same neuron as shown in the boxes. *fm*s in the boxes represent the content of documents in the collection.

Figure 2. The example of the document matrix at the given frequency threshold value *f* is equal to 2

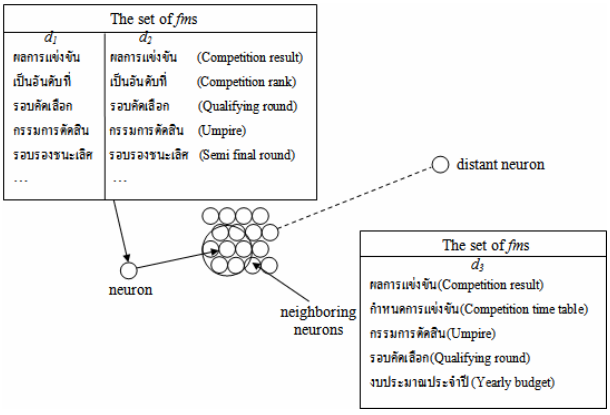

Figure 3. The example of the document cluster map

After the SOM has been trained, the document clusters are formed by labeling each neuron that contains certain documents of similar type. The documents in the same neuron may not contain exactly the same set of frequent max substrings or *FMAX*, but they usually contain mostly overlapping frequent max substrings. As a result, the document cluster map can be used as a prediction model to generate the different groups of similar documents, and each group will then be used to specify the document type by comparing frequent max substrings of each group with keywords of each area. In **Figure 4**, we depict clustering the documents into different groups, by mapping an input data with neurons in the document cluster map to find the document groups of several types.

From **Figure 4**, the following will describe the process of matching an input data x with the neurons in the document cluster map by using SOM.

Let us consider the input vector $x = [x_1, x_2, \ldots, x_n]^t \in R^n$ as the input data sets where t is the frequent max substrings of the input documents. These input data sets have to be matched with all neurons in the map that is denoted as two-dimensional network of cells or the model vector $m_i = [m_{i1}, m_{i2}, \ldots, m_{in}]^t \in R^n$ depicted in **Figure 5**. Each neuron i in the network contains the model vector m_i, which has the same number of indexing terms as the input vector x.

From **Figure 5**, the input vector x is compared with all neurons in the model vector m_i to find the best matching node called the winner. The winner unit is the neuron on the map where the set of the frequent max substrings of the input vector x is the same or similar to the set of the frequent max substrings of the model vector m_i by using some matching criterion e.g. the Euclidean distances between x and m_i. As a result, this method can be used to cluster documents into different groups, and also suggested that this can use to reduce the search time for the relevant document.

6. Experimental Studies and Comparison Results

In this section, we describe an experiment for clustering non-segmented documents (Thai documents) based on the proposed SOM and frequent max substring technique. We also compare the proposed technique with the SOM based documents clustering using single words as features and hierarchical clustering technique using single words on a group of documents. 50 Thai documents were used as an input dataset to train a map. All Thai documents used are from Thai news websites that consist of different categories of contents: sport, travel, education and political news. The documents have varying lengths. The set of documents contains 103,287 characters, and average document length is 2,065 characters or 78 words per document. The basic statistics for the text collection are shown in **Table 1**.

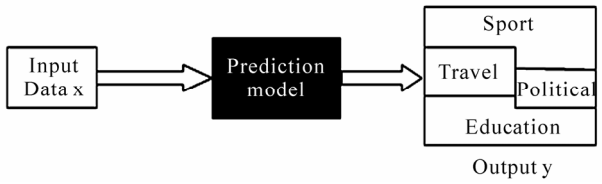

Figure 4. Neuron network architecture

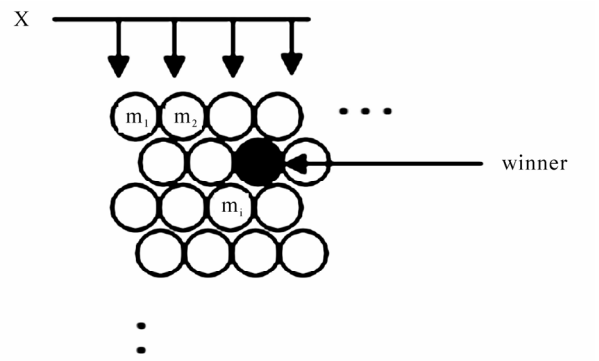

Figure 5. Self-organizing map

Table 1. Basic statistics for Thai text collection

	No. of Docs	No. of Chars	No. of Words	Avg. Chars./ Docs	Avg. Words/ Docs
Sport news	15	24727	997	1648.46	66.46
Travel news	15	29096	1022	2078.28	73
Political news	15	38017	1398	2534.46	93.2
Education news	5	9445	336	1889	67.2

In our proposed technique, the set of frequent max substring was first generated by frequent max substring technique at the given frequency threshold value, which is equal to 2 from the document dataset and 35 frequent max substrings, the long and frequently occurring terms in sport, travel, political and education documents, were used as the set of indexing terms for this document collection. The 50 input documents are then transformed to a document matrix of weighted frequent max substring occurrence. Hence, these 35 indexing terms and 50 input documents to form a 50 × 35 matrix, where each document vector was represented by each column of the matrix, and the rows of the matrix correspond to the indexing terms. We use this 50 × 35 matrix to train a map using SOM, and the number of neurons was set as 9 in SOM program as shown in **Figure 6**.

In the experimental study, 9 was set as the number of neurons because the several numbers of neurons were investigated, and 9 neurons provided the best result among them. From experimental studies, the group of political, sport, and education documents provided good results as

similar documents of each type were mapped onto the neuron. It can be observed that some errors occurred within the group of travel documents. The travel documents were mapped onto several neurons due to overlapping terms that appeared across different type of documents.

The **Figure 6** showed the map containing 9 neurons and 50 Thai documents. Each neuron contains the group of similar documents.

From **Figure 6**, the experimental result showed that SOM can cluster 50 documents into 5 neurons on the map, and the similar documents were grouped into the same neuron as shown in **Table 2**.

As observed from the results, this technique can be used to cluster non-segmented documents into several groups according to their similarity. The accuracy of this technique is up to 83.25%. However, from this experiment, we have found that the groups of education and sport documents are mapped onto the same neuron (Neuron5) because they both contain mostly overlapping frequent max substrings such as ผลการแข่งขัน (competition result), การจัดอันดับ (position ranking), ได้รับรางวัล (getting award), etc. Furthermore, the contents of documents and generated indexing terms are also the main factors that impact the accurate value. The content of one document may have overlapping terms from two different types of documents. For instance, Education5, Travel 1, Travel 3, Travel 11 and Travel 14 documents are mapped onto the neuron 3 because they are presenting information on ecotourism, containing overlapping terms from education and travel documents.

The methodology to measure the clustering approaches is to compare the group of documents occurrences. In this paper, we compare our technique with the SOM based documents clustering using single words [31] as features and hierarchical based document clustering using single words [4,12]. In the SOM based documents clustering using single words, single words were first extracted by using Thai word segmentation from the same document data set and 35 single words, most frequent occurring single keywords in education, sport, travel and political documents, and 50 input documents are used to form a 50 × 35 matrix, which is the same matrix size as our earlier experiment to train a map using SOM. From experimental studies, it can be observed that only the groups of politic, and travel documents provided fairly good results as shown in **Table 3**.

The group of similar travel documents was mapped onto the neuron as well as the group of political documents. Meanwhile, the education and sport documents were distributed onto several neurons because most single words extracted from education and sport documents are general terms. These general terms can be shared in many types of documents.

To depict the experiment results, **Figure 7** showed the map containing 9 neurons and 50 Thai documents, and the documents were grouped into the neurons as shown in **Table 3**.

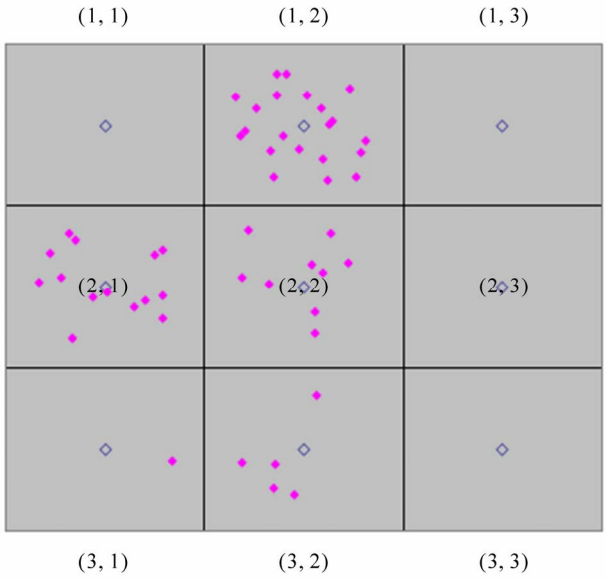

Figure 6. SOM contains 9 neurons and the group of similar documents from 50 Thai document collection

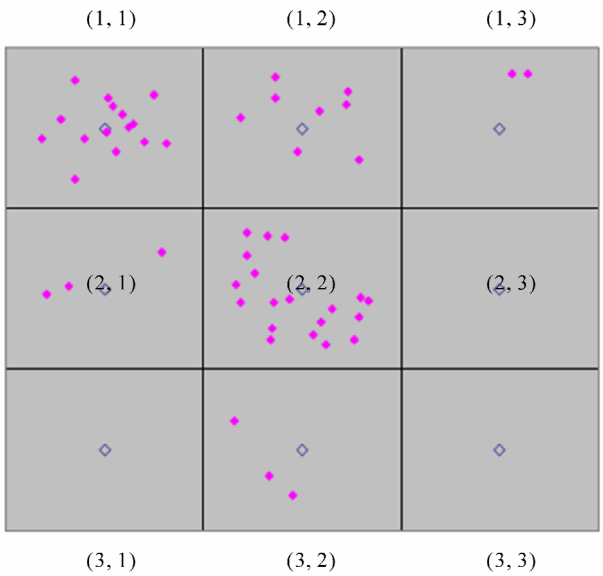

Figure 7. SOM contains 9 neurons and the group of similar documents from 50 Thai document collection by using single words as features

It can be observed that the accuracy of the SOM based documents clustering using single words is 72.21%. Moreover, from the experiment results, we have found that the SOM based documents clustering using frequent max substrings provides better result than the SOM based documents clustering using single words because the frequent max substrings can be used to describe the

Table 2. Clustering results of using SOM and frequent max substring technique

Neuron ID	Row	Column	Document ID
Neuron 5	1	2	Political 1, Education 1, Education 2, Education 3, Education 4, Sport 1, Sport 2, Sport 3, Sport 4, Sport 5, Sport 6, Sport 7, Sport 8, Sport 9, Sport 10, Sport 11, Sport 12, Sport 13, Sport 14, Sport 15, Travel 10
Neuron 2	2	1	Political 2, Political 3, Political 4, Political 5, Political 6, Political 7, Political 8, Political 9, Political 10, Political 11, Political 12, Political 13, Political 14, Political 15,
Neuron 4	2	2	Travel 2, Travel 4, Travel 5, Travel 6, Travel 7, Travel 8, Travel 9, Travel 13, Travel 15
Neuron 1	3	1	Travel 12
Neuron 3	3	2	Education 5, Travel 1, Travel 3, Travel 11, Travel 14

Table 3. Clustering results of SOM based documents clustering using single words

Neuron ID	Row	Column	Document ID
Neuron 2	1	1	Education 1, Sport 1, Sport 2, Sport 5, Sport 6, Political 1, Political 2, Political 3, Political 4, Political 6, Political 8, Political 9, Political 10, Political 11, Political 12
Neuron 5	1	2	Education 3, Education 5, Sport 7, Sport 8, Sport 9, Sport 11, Sport 13, Travel 7
Neuron 6	1	3	Sport 12, Sport 14,
Neuron 1	2	1	Sport 15, Travel 15, Politica l7
Neuron 4	2	2	Education 2, Sport 3, Sport 4, Travel 1, Travel 2, Travel 3, Travel4, Travel 5, Travel 6, Travel 8, Travel 9, Travel 10, Travel 11, Travel 12, Travel 13, Travel 14, Political 13, Political 14, Political 15
Neuron 3	3	2	Education 4, Sport 10, Politcal 5

content of the documents more specifically than using single words, as the frequent max substrings can be referred to the frequently and long terms rather than individual words. Moreover, many researches have also shown an improvement in clustering performance when using phrases rather than single words as features [21,22].

Additionally, we also compare our technique with the hierarchical based document clustering using single words. The hierarchical based document clustering is used because it has been widely used and has been applied successfully in many applications in the area of document clustering [12]. This method has also been used to perform Thai documents clustering. To use this technique with Thai language, single words were first extracted by using Thai word segmentation techniques from the same document dataset used in our experiment that is discussed earlier. After word segmentation is performed, single words are then transformed into feature vector of the words that appear in documents. The term-frequencies of the words are also contained in each feature vector. The feature vector of the words is then used to compute the similarity of the documents by using the hierarchical clustering approach. In the hierarchical clustering program, the feature vector of the words with their frequencies was used as an input data, and the number of clusters was set to 9 after trial-and-error. From the experimental results, the group of political, travel and education documents provided good results. However, travel and education documents were grouped into the same cluster (cluster 1). It can be observed that the travel and education documents are sharing overlapping words that appeared across different type of documents. In addition, some of travel, education, sport and travel document were distributed across several small clusters. In **Table 4**, the experimental result showed that the hierarchical clustering program can cluster 50 documents into 9 clusters.

As can be observed from the results, the accuracy of the proposed method is up to 83.25%, meanwhile using the SOM based documents clustering using single words and the hierarchical clustering approach provide the accuracies 72.21% and 79.75% respectively. The hierarchi-

Table 4. Clustering results of using the hierarchical clustering approach

Cluster ID	Document ID
Cluster 1	Education 1, Education 2, Education 3, Education 4, Sport 9, Sport 13, Travel 1, Travel 2, Travel 3, Travel 4, Travel 5, Travel 7, Travel 8, Travel 9, Travel 10, Travel 11, Travel 12, Travel 13, Travel 14, Travel 15, Political 15
Cluster 2	Education 5
Cluster 3	Sport 1
Cluster 4	Sport 2, Sport 3, Sport 4, Sport 5, Sport 6, Sport 7, Sport 8, Sport 15
Cluster 5	Sport 10, Sport 11
Cluster 6	Sport 12
Cluster 7	Sport 14
Cluster 8	Travel 6
Cluster 9	Political 1, Political 2, Political 3, Political 4, Political 5, Political 6, Political 7, Political 8, Political 9, Political 10, Political 11, Political 12, Political 13, Political 14

cal clustering approach also created many small clusters that containing only a few documents. As a result, an improvement was demonstrated using frequent max substrings rather than single words as features. This proposed technique also does not require any pre-processing technique to extract the frequent max substrings. Meanwhile, the SOM based documents clustering using single words and the hierarchical based document clustering using single words require word segmentation to extract the single words.

7. Conclusions

This paper describes a non-segmented document clustering method using self- organizing map (SOM) and frequent max substring technique to improve the efficiency of information retrieval. We first use the frequent max substring technique to discover patterns of interest, called frequent max substrings, rather than individual words from Thai text documents, and these frequent max substrings are then used as indexing terms with their number of occurrences to form a document vector. SOM is then applied to generate the document cluster map by using the document vector. The experiment studies and comparison results on clustering the 50 Thai text documents is presented in this paper. We compare the proposed technique with the SOM based documents clustering using single words and hierarchical based document clustering technique with the use of single words for grouping the document occurrences. From the experimental results, our technique can be used to cluster 50 Thai documents into different clusters with more accuracy than using the SOM based documents clustering using single words and the hierarchical clustering approaches. As a result, the generated document cluster map from our technique can be used to find the relevant documents according to a user's query more efficiency.

REFERENCES

[1] B. Liu, "Web Data Mining: Exploring Hyperlinks, Con-tents, and Usage Data," 1st Edition, Springer-Verlag, New York Berlin Heidelberg, 2007.

[2] D. R. K. R. D. Cutting, J. O. Pedersen, J. W. Tukey, "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections," *Proceedings of ACM Special Interest Group on Information Retrieval* '92, Copenhague, 1992, pp. 318-329.

[3] I. Matveeva, "Document Representation and Multilevel Measures of Document Similarity," *Irina Matveeva, Document representation and multilevel measures of document similarity, Proceedings of the* 2006 *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*: *companion volume*: *doctoral consortium*, New York, 2006, pp. 235-238.

[4] G. K. M. Steinbach and V. Kumar, "A Comparison of Docu-ment Clustering Techniques," *KDD Workshop on Text Mining*, Boston, 2000.

[5] A.-H. Tan, "Text Mining: The state of the art and the challenges," *Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases*, Beijing, 1999, pp. 65-70.

[6] Q. L. H. Jiao and H.-B. Jia, "Chinese Keyword Extraction Based on N-Gram and Word Co-occurrence, 2007 *International Conference on Computational Intelligence and Security Workshops* (*CISW* 2007), Harbin, 2007, pp. 124-127.

[7] J. Mathieu, "Adaptation of a Keyphrase Extractor for Japanese Text," *Proceedings of the* 27*th Annual Conference of the Canadian Association for Information Science* (*CAIS*-99), Sherbrooke, Quebec, 1999, pp. 182-189.

[8] T. Chumwatana, K. W. Wong and H. Xie "An Automatic Indexing Technique for Thai Texts Using Frequent Max Substring," 2009 *Eight International Symposium on Natural Language Processing*, Bangkok, 2009, pp. 67-72.

[9] R. Feldman and J. Sanger, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data," Cambridge University Press, Cambridge, 2006.

[10] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data," Prentice Hall, New Jersey, 1988.

[11] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," John Wiley and Sons, New York, 1990.

[12] G. K. Y. Zhao, "Comparison of Agglomerative and Partitional Document Clustering Algorithms," The SIAM workshop on Clustering High-dimensional Data and Its Applications, Washington, DC, April 2002.

[13] Z. Huang, "Extensions to the K-means Algorithm for Clustering Large Datasets with Categorical Values," *Data Mining and Knowledge Discovery*, Vol. 2, No. 3, 1998, pp. 283-304.

[14] D. Dembele and P. Kastner, "Fuzzy C-Means Method for Clustering Microarray Data," *Bioinformatics*, Vol. 19, No. 8, 2003, pp. 973-980.

[15] L. J. Heyer, S. Kruglyak and S. Yooseph, "Exploring Expression Data: Identification and Analysis of Coexpressed Genes," *Genome Research*, Vol. 9, No. 11, 1999, pp. 1106-1115.

[16] C. C. Fung, K. W. Wong, H. Eren, R. Charlebois and H. Crocker, "Modular Artificial Neural Network for Prediction of Petrophysical Properties from Well Log Data," *IEEE Transactions on Instrumentation & Measurement*, Vol. 46, No. 6, December 1997, pp. 1259-1263.

[17] D. Myers, K. W. Wong and C. C. Fung, "Self-organising Maps Use for Intelligent Data Analysis," *Australian Journal of Intelligent Information Processing Systems*, Vol. 6 No. 2, 2000, pp. 89-96.

[18] D. R. Hill, "A Vector Clustering Technique," In: Samuelson, Ed., *Mechanized Information Storage, Retrieval and Dissemination North-Holland*, Amsterdam, 1968.

[19] J. J. Rocchio, "Document Retrieval Systems — Optimization and Evaluation," Doctoral Thesis, Harvard University, Boston, 1966.

[20] A. W. G. Salton and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communication of ACM*, Vol. 18, No. 11, 1975, pp. 613-620.

[21] O. Zamir, "Clustering Web Documents: A Phrase-Based Method for Group Search Engine Results," *Computer Science & Engineering*, Ph.D. Thesis, University of Washington, 1999.

[22] M. F. H. J. Bakus and M. Kamel, "A SOM-Based Document Clustering Using Phrases," *Proceeding of the* 9*th International Conference on Neural Information Processing* (*ICONIP*'02), Vol. 5, 2002, pp. 2212-2216.

[23] D. Mladenic and M. Grobelnik, "Word Sequence as Features in Text-learning," *Proceedings of the* 17*th Electro technical and Computer Science Conference* (*ERK*-98) Ljubljana, Slovenia, 1998.

[24] K.-H. Tsai, C.-M. Tseng, C.-C. Hsu and H.-C. Chang, "On the Chinese Document Clustering Based on Dynamical Term Clustering," *Asia Information Retrieval Symposium* 2005, Jeju Island, October 2005, pp. 534-539.

[25] C. Kruengkrai and C. Jaruskulchai, "Thai Text Document Clustering Using Parallel Spherical K-means Algorithm on PI-RUN Linux Cluster (in Thai)," *The* 5*th National Computer Science and Engineering Conference*, Chiang Mai University, Chiang Mai, 2001, pp. 7-9

[26] T. Kohonen, "Self-Organization and Associative Memory," *Springer Series in Information Sciences*, Springer-Verlag, Berlin, 1984, p. 125.

[27] T. Chumwatana, K. W. Wong and H. Xie "Frequent max substring mining for indexing," *International Journal of Computer Science and System Analysis* (*IJCSSA*), India, 2008, pp. 179-184.

[28] T. Chumwatana, K. W. Wong and H. Xie "An Efficient Text Mining Technique," 9th *Postgraduate Electrical Engineering & Computing Symposium* (*PEECS*2008), Perth, Australia, 2008, pp. 147-152.

[29] T. Chumwatana, K. W. Wong and H. Xie, "Using Frequent Max Substring Technique for Thai Keyword Extraction used in Thai Text Mining," 2*nd International Conference on Soft Computing, Intelligent System and Information Technology* (*ICSIIT* 2010), Bali, 1-2 July 2010, pp. 309-314.

[30] T. Chumwatana, K. W. Wong and H. Xie, "Thai Text Mining to Support Web Search for E-Commerce," *The* 7*th International Conference on e-Business* 2008 (*INCEB* 2008), Bangkok, 2008, pp. 66-70.

[31] J. E. Hodges and Y. Wang, "Document Clustering using Compound Words," *Proceedings of the* 2005 *International Conference on Artificial Intelligence* (*ICAI* 2005), Las Vegas, Nevada, 2005, pp. 307-313.

