Documents Clustering using Partitional Clustering Methods

Khin Myo Wai ,Khin Cho San
University of Computer Studies, Magway
k.khinmyowai@gmail.com

Abstract

Document clustering is text processing that groups documents with similar concept. Clustering is defined as a process of partitioning a set of objects (patterns) into a set disjoined group (clusters). Its goal is to reduce the amount of data by categorizing or grouping similar data items together and obtain useful information. Clustering methods can be divided into two basic types: hierarchical and partitional clustering. This system used two partitional clustering methods. They are Self-Organizing Map (SOM) and K-Means. Self-Organization Maps is an artificial neural network model that is well suited for mapping high dimensional data into a two-dimensional representation space. SOM clustering is one of the well-known unsupervised clustering techniques. The goal of K-Means is to find k points of a dataset that can best represent the dataset in a certain mathematical sense (to be detailed later). These k points are also known as cluster centers, prototypes, centroids, or code words, and so on. The most known class of partitioned clustering algorithms is the K-Means algorithm and its variants. In this paper, documents are clustered by SOM algorithm how these are related to each other and K-Means start by randomly selecting k point cluster means; then assigns each document to its nearest cluster mean.

1. Introduction

The process of document clustering is collecting similar documents into clusters. Document clustering techniques have been receiving more and more attentions as a fundamental and enabling tool for efficient organization, navigation, retrieval and summarization of huge volumes of text document. The partitioning based clustering methods are SOM, K-Means methods, Fuzzy c-mean and frequent term set association methods. In this system, SOM and K-Means are used for document clustering. The documents with text file format are inputted to the system and the words in the files are processed by stop words removing and stemming as the preprocessing. And then the words are counted as the vector representation. The vector representations, keywords, are inputted to the clustering algorithms, SOM algorithm and K-Mean algorithm. These keywords are used to calculate for clustering. After clustering, the result of the SOM is the number of unknown cluster and the result of the K-Mean is the number of predefined cluster.

Documents clustering is widely applicable in areas such as search Engines, Information Retrieval and Web Mining. The commonly used analysis is to run a clustering process based on the similarity among these documents. Document is a cluster very likely to match the same information need, but evaluation of the cluster quality is still a major concern.

2. Clustering

Clustering is a very useful and important technique for analyzing data. Clustering is similar to the classification but there has the differences. Classification is supervised categorization when classes are known and clustering is unsupervised categorization when classes are not known. Both classification and clustering have benefit from the integration of prior external class knowledge, which reflects specific classification concepts or organization. Classification knowledge can be used in the clustering algorithm, which groups documents by the influence of domain knowledge[5].

2.1 Type of Clustering Methods

Different types of major clustering methods are

- Partitional methods,
- Hierarchical methods,
- Dendity-based methods,
- Grid-based methods and
- Model-based methods

This system based on Self-Organizing Map (SOM) and K-Means in partitional methods.

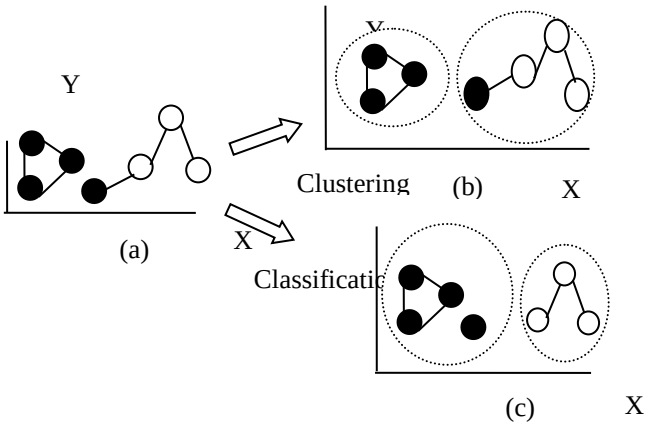

Figure 1: Example of Different Decision from Document Clustering and Human Classification

3. Documents Clustering

Document clustering is "unsupervised" learning in which there is no predefined category or class, but groups of documents that belong together are sought. Document clustering assists the retrieval by creating links between similar documents. The majority of documents clustering techniques fall into major categories. The majority of documents clustering techniques fall into two major categories. They are Hierarchical Clustering and Partitional Clustering methods [1].

Clustering Partitions a set of object into non-lapping subsets called clusters such that the object inside each cluster are similar to each other and the object from different clusters are not similar. The set of non-overlapping cluster is called a partition.

The text clustering task resembles the text categorization task. By definition, text categorization is the assignment of neural language text to one or more predefined categories, based on their contents [7].

Text categorization is a kind of "supervised" learning where the cat.gories are known beforehand and determined in advance for each training document. In contrast, document clustering is "unsupervised" learning in which there is no predefine category or "class" , but groups of documents that belong together are sought.

4. Self-Organizing Maps (SOM) and K-Means

4.1 Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM) is unsupervised algorithm. It allows multidimensional inputs to be mapped to a two or three dimensional map. A type of neural network that is particularly well-suited to clustering and visualization. It is a variation of prototype-based clustering. The goal of SOM is to find a set of centroids (reference vectors in SOM terminology) and to assign each object in the data set to the centroid that provides the best approximation of that object[6]. Figure 2 shows a simple comparison between a back-propagation neural network and a Self-Organizing Map neural network[2].

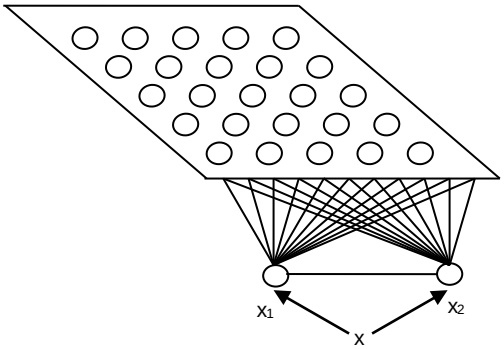

Figure 2: A Self-Organizing Network

4.1.1 SOM Algorithm

Use the Euclidean distance to determine the similarity of vectors.

The Euclidean distance, $\|x-y\|$, where $x = (a1,a2, \ldots ,an)$ and $y = (b1, b2, \ldots ,bn)$ is given by the equation:

Euclidean distance= $\sqrt{(a_1-b_1)2 + (a_2-b_2)2 + \ldots + (a_n-b_n)2}$

The general algorithm is as follows:

1. A set of input vectors is created, with each input vector consisting of ones and zeros, for example,
   [1 0 1 . . . . . 0 0 1].
2. Every node in the output map is represented by a reference vector with the same size as the input vectors. The values of the reference vectors are randomly initialized.
3. An input vector is chosen from the input set and compared to the reference vector for every node.
4. The reference vector that produces the smallest. Euclidean distance (or other metric) is considered the best match for that input vector.
5. The weights of the winning node and its neighbors are adjusted using the equation
   $w_i(t+1)=w_i(t) + \infty(t) (x-w_i(t))$
6. Return to step 2 and repeat the process for a predetermined number of iterations.

4.2 K-Means

K-means is an algorithm to classify or to group objects (documents) based on attributes or features into K number of group. The K-means algorithm takes the input parameter, k, that is, positive integer number. The grouping is done by minimizing the

sum of squares of distances between data and the corresponding cluster centroid [3].

4.2.1 K-Means Algorithm

Input: The number of clusters k and a database containing n objects.
Output: A set of k clusters that minimizes the square-error criterion;
Method:

1. arbitrary choose k objects as the initial clusters;
2. repeat
3. (re)assign each objects to the cluster to object is the most similar, based on the mean value of the objects in the cluster;
4. update the cluster means, i.e., calculate the mean value of the objects of the each cluster;
5. until no change;

5. System Design

The fundamental design of the system is show in figure 3. The system consists of input documents, preprocessing and vector representation .And then calculate the cluster result using Self-Organizing Map (SOM) and K-Means methods.

5.1 Input Documents

The input documents are text flies as keywords.The text files are create in the notepad.

5.2 Document Preprocessing

For the training of SOMs, the documents must be encoded in form of numerical vectors. To be suited for the learning process of the map, to similar documents with similar conents should be close to each other, and possibly assigned to the same neuron[1].The presented approach is based on statisitical evaluation of work occurrence. It is important to be able to calculate significant statistics. Therefore, the number of considered words must be kept reasonably small, and the occurrences of words or by grouping words with equal or similar meaning. A possible way to do so is to filter so-called stop words and build the stems of the words.

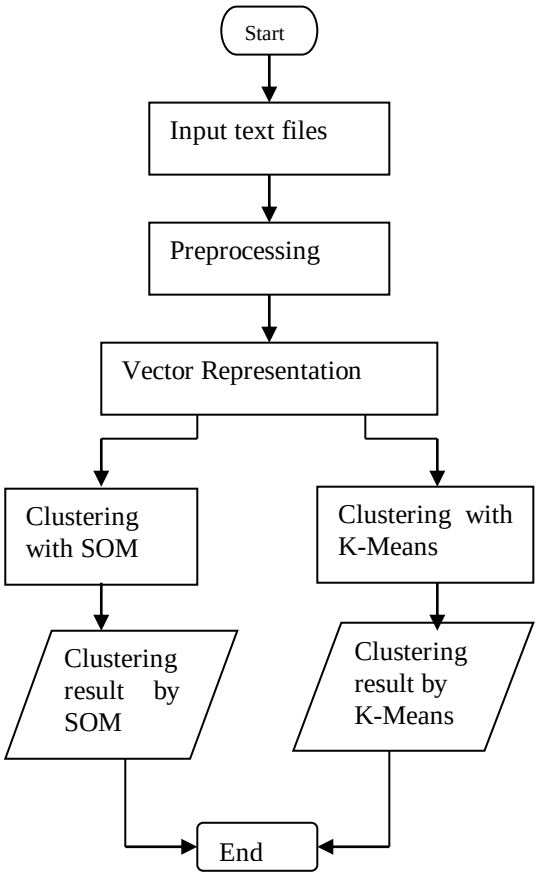

Figure 3: System Design

5.2.1 Stop Words Removing

In this system, abstract of raw documents are inputted to the system. Then the system remove the stop words and some example of these are given below.

Article : a, an, the, … , etc.
Subject : he, she, they, …, etc.
Preposition : in, for, out , under, …, etc.
Conjunction : and, or, that, after, … , etc.
Object : me, him, … , etc.

5.2.2 Stemming

In stemming, the similar meanings of words are grouped together. For example,

counti=country=countries
pagoda=pagodas
neighbour=neighbouring
divis=division=divisions
hundr=hundred

An axample of documents clustering,
The input documents of text file is the following,
File 1

My Country

My native land, Myanmar Naing-Ngan, is in Southeast Asia . It is also known as the land of Golden Pagodas. Its neighbouring countries are India, Chine Bangladesh, Laos and Thailand.

My country is made up of seven states and seven divisions. Its area is over two hundred and sixty thousand miles.

Preprocessing

- **Stop Word removing**
- **Stemming**

The following result is stop word removing function.

Country native land , Myanmar Naing-Ngan , Southeast Asia . also known land Golden Pagodas. neighbouring countries India ,Chine Bangladesh,Laos Thailand. country made up seven states seven divisions . area over two hundred sixty thousand miles.

And then, The stemming take the similsar meaning of words as a group and change capital letter to small letter. It delete comma(,), fullstop(.), full-comma(;), dash(-) etc.

The result of stemming is as following:

countri nativ land myanmar na ngan southeast asia also known land golden pagoda neighbour countri india chine bangladesh lao thailand countri made up seven state seven divis area over two hundr sixti thousand mile

The following file names are clustering text files for calculate SOM and K-Means results. These are twenty files.

File 1 = My Country
File 2 = Self-Organizing Map
File 3 = Grouping SOM
File 4 = My Faimily
File 5 = Application of SOM
File 6 = Hierarchical Clustering
File 7 = My Dear Sister
File 8 = Cluster with partition methods
File 9 = My Parents
File 10 = My Brother
File 11 = Clustre methods
File 12 = Document Clustering
File 13 = Classification
File 14 = My Dear Friends
File 15 = Clustering and SOM
File 16 = My house
File 17 = I visit to a city
File 18 = Unsupervised clustering methods
File 19 = Labeling Clusters
File 20 = My Class Room

Table 1: Vector Representation keywords

File name	countri	nativ	land	myanmar	na	ngan	souhtest	asia	also	…
File 1	3	1	2	1	1	1	1	1	1	…
File 2	0	0	1	0	1	2	3	0	0	…
…	…	…	…	…	…	…	…	…	…	…

For SOM, the number of cluster are not predefine as SOM is unknown cluster that produces the smallest. It is considered the best match for that input vector.

First iteration for SOM from table 1,

Use Euclidean distance

For eg; (File 4,File 5)=$\sqrt{}$(3-0)2+(1-0)2+…

=3.74

(File 4, File 6)=5.12

(File 4,File 7)=10.12

•
•
•

The weight of the winning node and its neighbours are adjusted using the equation for second input.

$w_i(t+1)=w_i(t) + \infty(t) (x-w_i(t))$

For eg; Let $\infty=0.5$, $w_i=1$

input(coutri)=1+0.5(3-1)=2

•
•
•

input(known)=1+0.5(1-1)=1

•
•
•

The reference vector that produce the smallest.

For K-Means, the number of cluster are predefine before the cluster. Because K-Mean is user define. Based on means value, this system calculate the cluster which the user gives. Although K-Means

cluster documents based on mean value, SOM do based on learning weight. As soon as the number of cluster change, mans value will change. In K-Means, there is a change in answer by the number of cluster. So, SOM can better give exact result than K-Means.
First iteration for K-Means from table 1,
The distance function is Eclidation distance
For eg; Let File 4 and File 5 as the center of each cluster.

Table 2: First iteration table for K-Means

File name	File 4 Distance Mean 1	File 5 Distance Mean 2	Cluster
File 1	0	3.74	1
File 2	3.74	0	2
File 3	3.45	2.11	2
. .	. .	. .	. .

To cluster the text files with K-Mean, the first must define the number of clusters is three.

K-Means Result is as Following :=

Cluster:(1)	Cluster:(2)	Cluster:(3)
File 1	File 2	File 6
File 4	File 3	File 8
File 7	File 5	File 11
File 9	File 15	File 12
File 10		File 13
File 14		File 16
File 17		File 18
File 20		File 19

SOM Result is as Following :=

Cluster(1)	Cluster(2)	Cluster(3)	Cluster(4)
File 4	File 2	File 6	File 1
File 7	File 3	File 8	File 17
File 9	File 5	File 11	File 20
File 10	File 15	File 12	
File 14		File 13	
File 16		File 18	
		File 19	

6. Conclusion

In this system, documents are clustered according to words in documents. After clustering the documents, each cluster containing similar related documents can be seen. These are calculated two methods with SOM (Self-Organizing Map) and K-Means. The clustering result of document collections has been presented by the integration of clustering methods. The user can see the result of document cluster as the user's criteria such as parameter value for the number of clusters to define the cluster integrity.

An additional issue related to selection an algorithm is correctly choosing the initial set of cluster. In the result, an adequate choice of clusters can strongly influence both the quality and the time required to obtain a solution. Some clustering methods are also important such as partitional clustering methods, need a distance matrix which contains all the distance between every pair of elements in the data set.

In this system, the different cluster between SOM and K-Means methods can be seen and the two main approaches to cluster the document of these methods can be compared.

5.1 Limitations

In order to make the conclusion more general, experiments with different documents collections are tested be necessary. In this system, text files can only be clustered. More over, does not permit over twenty files at a time and calculated accuracy with manual. This system uses the clustering methods of SOM and K-Means but does not use other clustering methods.

5.2 Further Extension

The user can view and study the clustering result of partitioning based clustering algorithms. This system can be extended using other Partitional Clustering or Hierarchical Clustering methods such as single linkage, complete linkage, overage linkage, ward's methods, Bisecting K-Means methods and hierarchical frequent termset association methods which should compare clustering methods. The system may cluster other files type such as doc files, word files, PDF files and etc. If calculate the accuracy, it would be better than these condition. The automatic organizing of documents by the SOM and K-Means combination should compare and calculate its precision and recall of the clustering results.

REFERENCE

1. Clustering Human Association
Raz Tamir, School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel imr@netvision.net.il

2. Gudio Deboeak, Ph.D*
"PUBLIC DOMAIN VERSUS COMMERCIAL TOOLS FOR EATING SELF-ORGANIZING MAPS'

3. Han Kamber "Data Mining" Concepts and Technique

4. Khaled M.Hammouda
"Web Documents A Preliminary Review"
Department of Systems Design Engineering, University of Waterloo,Waterloo,Ontario,Canada N2L 3G1,
hammouda@pami.uwaterloo.ca, Febrary 26,2001

5. Moses Charikar
"Clustering Methods"
Computer Science

6. Simon Haykin
" Neural Networks, A Computer- hensive foundation"
Second edition

7. http:// www.mathwork.com