International Journal of Computer and Communication Technology

Volume 2 | Issue 3 Article 1

July 2011

Soft clustering: An overview

A. B. Raut
Asst.Professor & Head, Department of Computer Science & Engg, HVPM's COET, Amravati , INDIA,
abrout@gmail.com

G. R. Bamnote
Asst.Professor & Head, Department of Computer Science & Engg., PRMITR, Badnera, INDIA,
grbamnote@gmail.com

Recommended Citation
Raut, A. B. and Bamnote, G. R. (2011) "Soft clustering: An overview," *International Journal of Computer and Communication Technology*: Vol. 2 : Iss. 3 , Article 1.
DOI: 10.47893/IJCCT.2011.1086
Available at: https://www.interscience.in/ijcct/vol2/iss3/1

Soft clustering: An overview

Prof. Ms. A. B. Raut
Asst.Professor & Head
Department of Computer Science & Engg.
HVPM's COET
Amravati , INDIA

Prof G. R. Bamnote
Asst.Professor & Head
Department of Computer Science & Engg.
PRMITR,
Badnera, INDIA

***Abstract*— Document clustering has been extensively investigated as a methodology for improving document retrieval process. In Traditional clustering algorithm each documents belongs to exactly one cluster & hence cannot detect the multiple themes of a document where as soft clustering algorithm each document can belong to multiple clusters. This paper gives a comparative study of hard clustering & soft clustering algorithm.**

Keywords- Document clustering, soft clustering, web mining

I. INTRODUCTION

The World Wide Web is a large repository of information. Over the last the decade we have witness an explosive growth in the information available on it . WWW become a major source of information.

Web mining is the use of data mining techniques to automatically discover & extract information from web documents & services[3] .Web mining can be decomposed into Resource discovery, information extraction and generalization subtasks. Web mining research can be classified into three categories : Web content mining (WCM) , Web structure mining(WSM), & Web usage mining (WUM)[1] .

Clustering is an important tool in web mining. The ability to automatically group similar icons together enables one to discover hidden similarity & key concepts . Clustering is also a powerful tool use to summaries a large amount of data into a small number of groups .

Finding the relevant information on www is not an easy task. The information user can encounter the following problems when interacting with the web[2].

1. low precision: Today's search tools have the low precision problem , which is due to the irrelevance of many search results. This results in a difficulty finding the relevant information .

2. Low recall: It is due to the inability to index all the information available on the web . This results in a difficulty finding the unindexed information that is relevant.

3. Creating new knowledge out of the information available on the web.: This problem could be regarded as a sub-problem of the problem above two problems .

4. Personalization of the information : This problem is often associated with the type & presentation of information , since it is likely that people differ in the contents & presentations they prefer while interacting with the web.

5. Learning about users : This is a problem that specifically deals with the problem 4 above , which is about knowing what the individual user's interests.

II. KEY REQUIREMENT FOR WEB DOCUMENT CLUSTERING

As pointed out by Zamir & Etzioni [5] the following are the key requirements for web document clustering methods .

1. Relevance : The method ought to produce that group documents relevant to the user's query .

2. Browsable summaries : The user needs to determine at a glance whether a cluster's contents are of interest . Ranked lists of the cluster may infact difficult to browse . Therefore the method has to provide concise & accurate descriptions of the clusters.

3. Overlap: since documents have multiple topics , it is important to avoid confining each document to only one cluster.

4. Snippet – Tolerance :The method ought to produce high quality cluster even when it only has access to the snippets returned by the search engines , as most users are unwilling to wait while the system downloads the original documents off the web.

5. Speed : A very patient user might sift through 100 documents in a ranked list presentation. Clustering on the other hand allows user to browse several related

documents . Therefore the clustering method ought to be able to cluster up to one thousand snippets in a few seconds. For the impatient user , each second counts.

6. Incrementality : To save time , the method should start to process each snippet as soon as it is received over the web.

III. RELATED WORK

Many Clustering algorithms have been developed & used in many fields[4,6]provides an extensive survey of various techniques . In this section we highlight work done on document clustering .

Broadly speaking there are two types of clustering algorithms: partition algorithms and hierarchical algorithms.

Agglomerative hierarchical clustering (AHC) algorithms are most commonly used It use a bottom –up methodology to merge smaller cluster into larger ones , using techniques such as minimal spanning tree . These algorithms find to be slow when applied to large document collection. It has different variants such as single-link, group-average and complete-link. Single-link and group-average methods typically takes $O(n^2)$ time while complete-link method typically takes $O(n^3)$ time.

Partition algorithm such as K- means are linear time algorithms . It try to divide data into subgroups such that the partition optimizes certain criteria , like inter – cluster distance or intra- cluster distances. They typically take an iterative approach. The time complexity of this algorithm is O(nkt) , where k is the number of desired clusters and T is the number of iterations. One advantage of K-means algorithm is that it gives overlapping clusters and the disadvantage is that it is most effective when the desired clusters are approximately spherical with respect to similarity measures used.

The modern clustering algorithms falls into four groups: partition algorithms, hierarchical algorithms, model fitting and density based. Model fitting algorithms such as COBWEB attempt to fit the data as a mixture of easily parameterized distributions. Density – based algorithm, such as DBSCAN view clustering as locating high-density regions.

Most of the early work applied traditional clustering algorithms like K-means to the set of documents to be clustered . Willett [4] provided a survey on applying hierarchical clustering algorithm into clustering documents. Cutting et al.[10] adapted various partition – based clustering algorithms to clustering documents . Two of the techniques are Buckshot & Fractionation . Buckshot selects a small sample of documents to pre-cluster them using a standard clustering algorithm & assigns the rest of the documents to the clustered formed. Fractionation splits N documents into ‘ m ‘ buckets where each bucket contains N/m documents . Fractionation takes an input parameter p, which indicates the reduction factor for each bucket . The standard clustering algorithm is applied so that if there are ‘ n ‘ documents in each bucket , they are clustered into n/p cluster . Now each of these cluster is treated as if they were individual documents & the whole process is repeated there are only ‘ K ‘ clusters.

Most of the algorithms consider the documents as bag of words(BOW) They use word base approach to find the similarity between two documents. A relatively new technique is proposed by Zamir & Etzioni [5]. They introduce the notion of phrase based approach for document clustering. They use a generalized phrase- tree to obtain information about the phases & use them to cluster documents.

IV. SOFT CLUSTERING ALGORITHMS

A single document often ally contains multiple themes like a web document on topic web mining may contain different themes like data mining, clustering and information retrieval. Many traditional clustering algorithms assign each document to a single cluster, thus making it difficult for the user to retrieve information.

Based on this concept we can divide clustering algorithm in to hard & soft clustering algorithm. In traditional clustering algorithm each documents belongs to exactly one cluster & hence can not detect the multiple themes of a document where as soft clustering algorithm each document can belong to multiple clusters .

Soft clustering has the following advantages:

- A document can belong to multiple clusters, thus user can find multiple themes for a document.
- Different clusters get formed for different themes.
- The measure associated between clusters and documents can be used as a relevance measure to order the document appropriately.

Fuzzy C-means(FCM) was given by Dunn .FCM is based on the partition clustering algorithm, iterating over the data set until the values of membership function stabilizes. World Base Soft Clustering (WBSC) approach is given by King – Ip Lin, Ravikumar[7]. It first forms initial clusters of the documents , with each cluster representing a single word for instance , WBSC forms a cluster for the word ‘ tiger’ made up of all the documents that contain the word ‘ tiger’ . After that , WBSC merges similar clusters – Clusters are similar if they contain the similar set of documents – using a hierarchical based approach until some stopping criterion is reached. At the end , the clusters are displayed based on the words associated with them. It consists of 3 steps that is cluster initialization , Cluster building & Display the result .

SISC (Similarity Based Soft Clustering algorithm) is a soft clustering algorithm [8]. It uses modified fuzzy C Means algorithm to cluster set of documents based on a given similarity measures. It use a randomization approach that

enables it to avoid a lot of computation needed in a traditional fuzzy clustering algorithms. At each iteration, it computes a similarity measure between a cluster and a document with a probability proportional to the proximity of the similarity measure to the threshold measure. SISC can broadly divided into four steps : A preprocessing steps to cleanup & transfer a data , a initial cluster generation steps to initialize the cluster ,an iterative step to build the cluster & post processing steps to present the result .

KMART algorithm is given by Kondadaldi and Kozma[9]. It is modified version of Fuzzy ART algorithm to employ soft clustering. This algorithm is broadly divided into stages i.e. pre-processing ,cluster building and keyword selection. The main advantage of KMART over most of fuzzy clustering algorithm is that the number of clusters is decided dynamically.

V. CONCLUSIONS AND FUTURE WORK

Soft clustering is today's need. Many soft clustering algorithms implements the idea of fuzziness in their method. Our future work involves to make these soft clustering algorithms more efficient by using better data structures.

REFERENCES:

[1] WangBin , LiuZhijing ,Web Mining Research , in proceeding of the 5th international conference on computational intelligence and multimedia applications(ICCIMA'03) 2003.

[2] R. Kosala,H. Blockeel.Web Mining Research: A survey,SIGKDD Explorations,ACM SIGKDD,July 2000.

[3] O.Etzioni,The World wide web:quagmire or gold mine. Comm. Of ACM, Nov 96.

[4] P. Willet, Recent trends in hierchical document clustering: a critical review, Information processing and management,1988

[5] O. Zamir,O. Etzioni,Web document clustering: a feasibility demonstration, in proceeding of 19th international ACM SIGIR conference on research and development in informational retrieval,1998.

[6] A. K. Jain,M. N. Murty and Flynn,Data clustering: A review, ACM computing surveys,1999.

[7] King-Ip Lin,Ravikumar Kondadadi,A word based soft clustering algorithm for documents, in proceeding of the 16th international conference on computers and their applications,(CATA-2001),March2001.

[8] King-Ip Lin, Ravikumar Kondadadi,A similarity based soft clustering algorithm for documents, in proceeding of the 7th international conference on database systems for advanced applications (DASFAA-2001),April 2001.

[9] Ravikumar Kondadadi and Robert Kozma, A modified fuzzy ART document clustering, IEEE transaction 2002.

[10] D. R. Cutting,D. R. Karger,J. O. Pedersen and J. W. Tukey, Scatter/Gather:a cluster-based approach to browsing large document collection, in proceeding of the 15th international ACM SIGIR conference on Research and development in information retrival,1992.