Volume 1, No. 2, July-August 2010

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Fuzzy Clustering in Web Mining

Ms.Anjali B. Raut *
Department of Computer Science & Engg.
HVPM's COET
Amravati, India.
anjali_dahake@rediffmail.com

Dr. G. R. Bamnote
Department of Computer Science & Engg.
PRMITR ,Badnera
Amravati,India.
grbamnote@rediffmail.com

Abstract: Conventional clustering means classifying the given data objects as exclusive subsets (clusters).That means we can discriminate clearly whether an object belong to a cluster or not. However such a partition is insufficient to represent many real situations. Therefore a fuzzy clustering method is offered to construct clusters with uncertain boundaries and allows that one object belongs to overlapping clusters with some membership degree. In other words, the essence of fuzzy clustering is to consider not only the belonging status to the clusters, but also to consider to what degree do the object belong to the cluster.

Keywords: Web Mining, Data Mining, Clustering, Hard Partition,Soft Partition,Information Retrieval

I. INTRODUCTION

The ability to form meaningful groups of objects is one of the most fundamental modes of intelligence. Human perform this task with remarkable ease. In early childhood one learns to distinguish, for example, between cats and dogs or apples and oranges. However, enabling the computer to do this task of grouping is a difficult and ill-posed problem.

Cluster analysis is a tool for exploring the structure of data. The core of cluster analysis is clustering; the process of grouping objects into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar. The need to structure and learn vigorously growing amount of data has been a driving force for making clustering a highly active research area.

Over the last decade there is tremendous growth of information on World Wide Web (WWW).It has become a major source of information. Netcraft survey report suggests that total count of websites is around 216 million (June 2010). Web creates the new challenges of information retrieval as the amount of information on the web and number of users using web growing rapidly.

Web Mining is the use of Data Mining techniques to automatically discover and extract information from web. Clustering is one of the possible techniques to improve the efficiency in information finding process. It is a Data Mining tool to use for grouping objects into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar. Web Mining has fuzzy characteristics, so fuzzy clustering is sometimes better suitable for Web Mining in comparison with conventional clustering.

II. RELATED WORK

Data Mining has emerged as a new discipline in world of increasingly massive datasets. Data Mining is the process of extracting or mining knowledge from data. Data Mining is becoming an increasingly important tool to transform data into information. Knowledge Discovery from Data i.e. KDD is synonym for Data Mining.

A. Web Mining

World Wide Web is a major source of information and it creates new challenges of information retrieval as the amount of information on the web increasing exponentially. Web Mining is use of Data Mining techniques to automatically discover and extract information from web documents and services [1].

Oren Etzioni was the person who coined the term Web Mining first time. Initially two different approaches were taken for defining Web Mining. First was a "process-centric view", which defined Web Mining as a sequence of different processes [1] whereas, second was a "data-centric view" , which defined Web Mining in terms of the type of data that was being used in the mining process [2]. The second definition has become more acceptable, as is evident from the approach adopted in most research papers[3][5]. Web Mining is also a cross point of database, information retrieval and artificial intelligence [4].

B. Web Mining Process

Web mining may be decomposed into the following subtasks:

1. Resource Discovery: process of retrieving the web resources.
2. Information Pre-processing : is the transform process of the result of resource discovery
3. Information Extraction: automatically extracting specific information from newly discovered Web resources.
4. Generalization: uncovering general patterns at individual Web sites and across multiple sites[3].

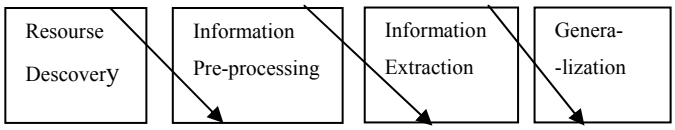

Figure I. Web mining Process

© 2010, IJARCS All Rights Reserves

C. Web Mining Taxonomy

Web has different facets that yield different approaches for the mining process:

1. Web pages consist of text.
2. Web pages are linked via hyperlinks
3. User activity can be monitored via Web server logs .

This three facets leads to the distinction into three categories i.e. Web content mining, Web structure mining and Web usage mining [4-7]. Following Fig 3 shows the Web Mining Taxonomy.

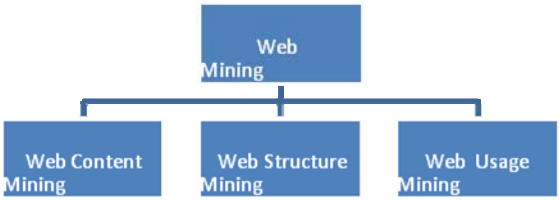

Figure II Web mining Taxonomy

a) Web Content Mining (WCM):

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched.

b) Web Structure Mining (WSM):

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web.

c) Web Usage Mining (WUM):

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data. Web usage data includes data from web server logs, browser logs, user profiles, registration data, cookies etc.

WCM and WSM uses real or primary data on the web whereas WUM mines the secondary data derived from the interaction of the users while interacting with the web.

D. Clustering

The Web is the largest information repository in the history of mankind. 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].

- 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.
- 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.

Clustering is one of the Data Mining techniques to improve the efficiency in information finding process. Many clustering algorithms have been developed and used in many fields. A. K. Jain, M. N. Murty and P. J. Flynn[8] provides an extensive survey of various data clustering techniques. Clustering algorithms can be broadly categorized into partitional and hierarchical techniques.

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. .

Most of the document clustering algorithm worked on BOW (Bag Of Words)model[5].Oren Zamir and Oren Etzioni[9] in their research listed the key requirements of web document clustering methods as relevance, browsable summaries, overlap, snippet tolerance, speed and accuracy. They have given STC (Suffix Tree Clustering) algorithm which creates clusters based on phrase shared between documents. Michael Steinbach, George Karypis and Vipin Kumar[10] presented the result of an experimental study of some common document clustering algorithms. They compare the two main approaches of document clustering i.e. agglomerative hierarchical clustering and K-means method. Nicholas O. Andrews and Edward A. Fox[11] presented the recent developments in document clustering . A single object oftenly 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 clustering algorithm can be divided into hard & soft clustering algorithm. In traditional clustering algorithm each object belongs to exactly one cluster where as in soft clustering algorithm each object can belongs to multiple clusters [12].

The conventional clustering algorithms in Data Mining have difficulties in handling the challenges posed by the collection of natural data which is often vague and uncertain. The modeling of imprecise and qualitative knowledge, as well as handling of uncertainty at various stages is possible through the use of fuzzy sets. Therefore a fuzzy clustering method was offered to construct clusters with uncertain boundaries, so this method allows that one object belongs to multiple clusters with some membership degree.

Pawan Lingras , Rui Yan and Chad West[13] applied fuzzy technique to discover usage pattern from web data. The fuzzy c-means clustering was applied to the web visitors of educational websites. The analysis shows the ability of the fuzzy c-means clustering to distinguish different user characteristics. Anupam Joshi and Raghu Krishnapuram[14] developed a prototype Web Mining system which analyzes web access logs from a server and tries to mine typical user access pattern. Maofu Liu, Yanxiang He and Huijun Hu[15] proposed a web fuzzy clustering model. In their paper the experimental result of web fuzzy clustering in web user

© 2010, IJARCS All Rights Reserves

clustering proves the feasibility of web fuzzy clustering in web usage mining.

III. FUZZY CLUSTERING ALGORITHMS

This section presents concept of hard and soft partition, Hard c–means (HCM) and Fuzzy c-means (FCM) algorithms mainly based on the descriptions in [16.18]. All algorithms described here are based on objective functions, which are mathematical criteria that quantify the quality of cluster models. The goal of each clustering algorithm is the minimization of its objective function.

A. Hard Partition

Let X be a set of data and x_i be an element of X. A Partition $p=\{C_1,C_2,\ldots,C_c\}$ of X is hard if and only if

i) $\forall x_i \in X \quad \exists C_j \in P$ Such that $x_i \in C_j$

/

ii) $\forall x_i \in X \quad x_i \in C_j \Rightarrow x_i \in C_j$ where $k \neq j$, $C_k, C_j \in P$

The first condition in the definition assures that the partition covers all data points in X,and the second condition assures that all clusters in the partition are mutually exclusive[16].

B Soft Partition

In many real world clustering applications , however , some data points partially belong to multiple clusters , rather than to a single cluster exclusively .For example a particular customer may be 'borderline case' between two groups of customers. These observations motivated the devalopment of the "soft clustering " algorithm.

A soft clustering algorithm find a soft partition of a given dataset based on certain criteria. In a soft partition , a datum can partially belong to multiple clusters.

Let X be a set of data and x_i be an element of X. A Partition $p=\{C_1,C_2,\ldots,C_c\}$ of X is soft if and only if and only if the following two conditions hold:

i) $\forall X_i \in X \quad \exists C_j \in P \quad 0 \leq \mu_{cj}(x_i) \leq 1$:

ii) $\forall X_i \in X \quad \exists C_j \in P$ such that $\mu_{cj}(x_i) > 0$.

Where $\mu c_j(x_i)$ denotes the degree to which x_i belongs to cluster C_j.

A type of soft clustering of special interest is one that ensures the membership degree of a point x in all clusters adding up to one , i.e.,

$$\sum \mu_{cj}(x_i)=1 \quad \forall x_i \in X$$

A soft partition that satisfies this additional condition is called a constrained soft partition. The fuzzy c-means algorithm, which is best known clustering algorithm, produces a constrained soft partition[16].

The following syntax will be used in the equations and algorithms :

J ..... objective function

$X = \{x_1,\ldots, x_n\}$ ... dataset of all objects (data instances)

$C = \{c_1,\ldots, c_c\}$ ... set of cluster prototypes (centroid vectors)

$D_{ij} = \|x_1 - c_1\|$... distance between object x_j and centre c_i

U_{ij} ... weight of assignment of object x_j to cluster i

C. Hard c – means (HCM)

Hard c-means is better known as k-means and in general it is not a fuzzy algorithm. However, its overall structure is the basis for all the others methods. Therefore it is called as hard c–means(HCM) in order to emphasize that it serves as a staring point for the fuzzy extensions.

The objective function of HCM can be written as follows:

$$J_h = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}\, d_{ij}^2 \qquad (3.1)$$

As HCM is a crisp algorithm, therefore:

$u_{ij} \in \{0,1\}$ .It is also required that each object belongs to exactly one cluster : $\sum_i (i = 1)^c u_{ij} = 1, \quad (j \in \{1,\ldots,n\})$.

The new membership weights is calculated as follows :

$$U_{ij} = \begin{cases} 1, \text{if } i = \operatorname{argmin}_{i=1}^{c} dlj, \\ 0, \quad \text{otherwise} \end{cases}$$

(3.2)

based on the weights , new cluster centres are calculated as:

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}\, x_j}{\sum_{j=1}^{n} u_{ij}} \qquad (3.3)$$

The algorithm can be stated very simple as shown below

Algorithm (3.1)The hard c-means(HCM) algorithm:

Randomly generate clusters centres

Repeat

For each object recalculate membership weights

using equation (3.2)

recompute the new centres using equation (3.3)

Until no change in c can be observed

The HCM algorithm has a tendency to get stuck in a local minimum, which makes it necessary to conduct several runs of the algorithm with different initializations. Then the best result out of many clustering can be chosen based on the objective function value.

© 2010, IJARCS All Rights Reserves

D Fuzzy c-means (FCM)

Probabilistic fuzzy cluster analysis [16] relaxes the requirement: $u_{ij} \in \{0,1\}$, which now becomes $u_{ij} \in [0,1]$.

However $\sum_i (i = 1)^c \square \square u \square_i j \square = 1, \quad (j \in \{1, \ldots, n\})$ condition still holds. FCM optimizes the following objective function :

$$J_f = \sum_{i=1}^{c} \sum_{j=1}^{n} \square u_{ij}^{m} d_{ij}^{2} \qquad (3.4)$$

Parameter m, m>1 , is called the fuzzyfier or the weighting exponent. The actual value of m determines the 'fuzziness' of the classification. It has been shown[ ] that for the case m=1 , J_f becomes identical to J_h & thus FCM becomes identical to hard c-means.

The transformation from the hard c-means to the FCM is very straightforward ;we must just change the equation for calculating memberships (3.2) with :

$$u_{ij} = \frac{1}{\sum_{i=1}^{c} \left(\frac{d_{ij}^2}{d_{ij}^2} \right)^{\frac{1}{m-1}}} = \frac{d_{ij}^{\frac{-2}{m-1}}}{\sum_{i=1}^{c} d_{ij}^{\frac{-2}{m-1}}} , \qquad (3.5)$$

and function for computing cluster centres (2.3) with :

$$c_i = \frac{\sum_{j=1}^{n} \square u_{ij}^{m} X_j}{\sum_{j=1}^{n} u_{ij}^{m}} . \qquad (3.6)$$

Equation (3.5) clearly shows the relative character of the probabilistic membership degree .It depends not only on the distance of the object X_j for the cluster C_i, but also on the distance between this object & other clusters . The algorithm can be stated as shown in Table 3.2

Algorithm {he Fuzzy c-means (FCM) algorithm} :

m: the parameter in the objective function

Є : a threshold for the convergence criteria

Intialize protype $V=\{v_1, v_2, \ldots, v_c\}$

Repeat

$V^{previous} \leftarrow V$

Compute membership functions using equation (3.5)

Recompute the new centers using equation (3.6)

Until $\sum || v_i^{previous} - v_i || \leq Є$

Although the algorithm(3.2) says the same as the algorithm(3.1). We get probabilistic output if we apply above changes. The (probabilistic) fuzzy c-means algorithm is known as a stable & robust classification method. Compared with the hard c-means it is quite insensitive to its initialization & it is not likely to get stuck in an undesired local minimum of its objective function in practice .Due to its simplicity & low computational demands , the probabilistic FCM is a widely used initializer for other more sophisticated clustering methods.

IV. CONCLUSION

In this paper we presents the concept of Web Mining and an overview HCM and FCM clustering algorithmsused in Web Mining. Web data has fuzzy characteristics, so fuzzy clustering is sometimes better suitable for Web Mining in comparison with conventional clustering.

Future work will consider the designing and implementing some techniques for converting fuzzy clusters to its crisp equivalent.

V. REFERENCES

[1] Oren Etzioni, "The World Wide Web: quagmire or gold mine?" ,Communications of ACM", Nov 96.

[2] R. Cooley,B. Mobasher and J. Srivastava ,"Web Mining: Information and Pattern Discovery on the World Wide Web", In the Proceeding of ninth IEEE International Conference on Tools with Artificial Intelligence(ICTAI'97),1997.

[3] Hillol Kargupta, Anupam Joshi, Krishnamoorthy Sivakumar and Yelena Yesha, "Data Mining: Next Generation Challenges and Future Directions", MIT Press,USA , 2004 .

[4] WangBin and LiuZhijing , "Web Mining Research" , In Proceeding of the 5th International Conference on Computational Intelligence and Multimedia Applications(ICCIMA'03) 2003.

[5] R. Kosala and H.Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations ACM SIGKDD, July 2000.

[6] Sankar K. Pal,Varun Talwar and Pabitra Mitra , "Web Mining in Soft Computing Framework : Relevance, State of the Art and Future Directions ", IEEE Transactions on Neural Network , Vol 13,No 5,Sept 2002 .

[7] Andreas Hotho and Gerd Stumme, "Mining the World Wide Web- Methods, Application and Perceptivities", in Künstliche Intelligenz, July 2007. (Available at http://kobra.bibliothek.uni-kassel.de/)

[8] A. K. Jain,M. N. Murty and P. J. Flynn, "Data clustering: A review," ACM computing surveys,31(3):264-323,Sept 1999.

[9] O. Zamir and O. Etzioni, "Web document clustering: A feasibility demonstration", in Proceeding of 19th International ACM SIGIR Conference on Research and Development in Informational Retrieval , June1998.

[10] Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document Clustering Techniques", KDD Worksop on Textmining, 2000.

[11] Nicholas O. Andrews and Edward A. Fox, "Recent Development in Document Clustering Techniques", Dept of Computer Science, Virgina Tech 2007.

[12] King-Ip Lin and 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.

[13] Pawan Lingras ,Rui Yan and Chad West, " Fuzzy C-Means Clustering of Web Users for Educational Sites", Springer Publication ,2003.

[14] Anupam Joshi and Raghu Krishnapuram, " Robust Fuzzy Clustering Methods to Support Web Mining", Proceedings of the Workshop on Data Mining and Knowledge Discovery , SOGMOD ,1998 .

[15] Maofu Liu, Yanxiang He and Huijun Hu ," Web Fuzzy Clustering and Its Applications In Web Usage Mining", Proceedings of 8th International Symposium on Future Software Technology (ISFST-2004).

© 2010, IJARCS All Rights Reserves

[16] John Yen and Reza Langari, "Fuzzy Logic : Intelligence,Control, and Information", Pearson Education,4th Edition , 2005 .

[17] Valente de Oliveira, J Pedrycz, " Advances in Fuzzy Clustering and its Application", John Wiley & Sons ,2000.

[18] Matjaz Jursic and Nada Lavrac ,"Fuzzy Clustering of Documents" In Slovenian KDD Conference of Data Warehouse , SIKDD'08.

© 2010, IJARCS All Rights Reserves