ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 1, April 2013

Comparison of Conventional & Fuzzy Clustering Techniques: A Survey

Neha Mehta[1], Mamta Kathuria[2,] Mahesh Singh[3]

Student, M.tech (CSE), Advanced Institute of Technology & Management, Palwal, Haryana, India[1]

Asstt. Proff.CSE, YMCA University of Science & Technology, Faridabad, Haryana, India[2]

Asstt. Proff.(CSE)Advanced Institute of Technology & Management, Palwal, Haryana, India[3]

Abstract: The web is the largest information repository observed till date. Due to its huge size however, finding the relevant information is not an easy task. So different searching and web mining techniques are being employed by the present day search engine for the purpose of information retrieval from the web. Web document clustering is one possible technique to improve the efficiency in information finding process. The traditional web mining, techniques of web mining have difficulties in handling the challenges posed by the collection of data which is vague and uncertain. Fuzzy clustering methods have the potential to manage such type of situations efficiently. This paper summarizes the different characteristics of web data, the web mining basics and limitations of existing web mining methods. The application of use of Fuzzy logic with web mining is being discussed with a view to highlight its importance in information retrieval. A comparative study of different fuzzy clustering techniques with the conventional clustering technique has been discussed.

Keywords*:* Web mining, Conventional clustering, Fuzzy Clustering.

I. INTRODUCTION

An explosive growth of information is available on the World Wide Web (WWW). Today, web browsers provide easy access to sources of text and multimedia data. More than 10 billion pages are indexed by search engines, and finding the desired information is not an easy task. This profusion of resources has prompted the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "web mining"[1].

Information Retrieval (IR) is the process of finding the material (usually documents) of unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers. IR [2] is used to find, extract, filter and order the desired information. IR Deals with automatic retrieval of all relevant documents. All non-relevant documents are fetched as few as possible. The field of information retrieval also covers supporting users in browsing or filtering document collections or further processing a set of retrieved documents. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents. It is similar to arranging books on a bookshelf according to their topic. Given a set of topics, standing information needs, or other categories (such as suitability of texts for different age groups), classification is the task of deciding which classes, if any, each of a set of documents belongs to. It is often approached by first manually classifying some documents and then hoping to be able to classify new documents automatically. Information retrieval systems can also be distinguished by the scale at which they operate, and it is useful to distinguish three prominent scales. In *web search*, the system has to provide search over billions of documents stored on millions of computers. Distinctive issues need to gather documents for indexing, being able to build systems that work efficiently at this enormous scale, and handling particular aspects of the web, such as the exploitation of hypertext and not being fooled by site providers manipulating page content in an attempt to boost their search engine rankings, given the commercial importance of the web.

To proceed toward web intelligence, obviating the need for human intervention, we need to incorporate and embed artificial intelligence into web tools. The necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge both across the internet and in particular web localities is drawing the attention of researchers from the domains of information retrieval, knowledge discovery, machine learning, and artificial intelligence (AI), among others. However, the problem of developing automated tools in order to find, extract, filter, and evaluate the users desired information from unlabeled, distributed, and heterogeneous web data is far from being solved. To handle these characteristics and here we use soft web mining to overcome some of the limitations of existing methodologies.

Soft computing [9] seems to be a good candidate for achieving useful information from the web; the research area combining the two may be termed as "soft web mining." Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another, flexible information processing capability for handling real-life ambiguous situations. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, low-cost solutions, and close resemblance to human-like decision making.

At present, the principal soft computing tools include fuzzy sets [4,5], artificial neural networks (ANNs), genetic algorithms (GAs), and rough set (RS) theory. Fuzzy sets provide a natural framework for the process in dealing with uncertainty. Neural networks (NNs) are widely used for modeling learning and adaption, genetic algorithms are used for optimization, and rough set theory used for handling uncertainty arising from limited discernibility of objects.

The rest of this paper is organized as follows: Section 2 deals with the characteristics of web data, and the different components and types of web mining. The limitations of existing web mining methods are discussed in Section 3. Section 4 provides an introduction about fuzzy logic and the importance of fuzzy clustering. Section 5 provides the conclusion and scope of future research in the area of soft web mining.

II. BASICS OF WEB MINING

A term coined in analogy to "data mining" referring to devising new techniques for classifying and extracting the useful information from the web. Web Mining is different from data mining because of the following:

A Characteristics of Web Data

The web is a vast collection of completely uncontrolled heterogeneous documents. Thus, it is huge, diverse, and dynamic, and raises the issues of scalability, heterogeneity, and dynamism, respectively. Due to these characteristics, we are currently drowning in information, but starving for knowledge; thereby making the web a fertile area of data mining research with the huge amount of information available online. Data mining refers to the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Web mining [1] can be broadly defined as the discovery and analysis of useful information from the WWW. In web mining data can be collected at the server side, client side, proxy servers, or obtained from an organization's database. Depending on the location of the source, the type of collected data differs. It also has extreme variation both in its content (e.g., text, image, audio, symbolic) and meta information, that might be available. This makes the techniques to be used for a particular task in web mining widely varying. Some of the characteristics of web data are

i) Unlabeled;
ii) Distributed;
iii) Heterogeneous (mixed media);
iv) Semi structured;
v) Time varying;
vi) High dimensional.

Therefore, web mining basically deals with mining large and hyperlinked information base having the aforesaid characteristics. Also, being an interactive medium, human interface is a key component of most web applications. Some of the issues which have come to light, as a result, concern

i) Need for handling context sensitive and imprecise queries;
ii) Need for summarization and deduction;
iii) Need for handling overlapped data.
iv) Need for handling imprecise data.

Thus, web mining, though considered to be a particular application of data mining, warrants a separate field of research, mainly because of the aforesaid characteristics of the data and human related issues.

2.2. Web Mining Categories

Web Mining is an application of "Data Mining" to extract large amount of information from the web. Web mining is divided into three categories namely web content mining (WCM), web structure mining (WSM), and web usage mining (WUM). The details are given in following sub sections.

2.2.1. Web Content Mining (WCM)

WCM [8] deals with the discovery of useful information from the web contents/data/documents/services. However, web contents are not only text, but encompass a very broad range of data such as audio, video, symbolic, metadata, and hyperlinked data. Out of these, research at present is mostly centered on text and hypertext contents. The web text data can be of three types:

1) Unstructured data such as free text;
2) Semi structured data such as HTML;
3) Fully structured data such as in tables or databases.

B. Web Structure Mining (WSM)

WSM [8] pertains to mining the structure of hyperlinks within the web itself (inter document structure unlike WCM, which pertains to intra document structure). Here, structure represents the graph of the links in a site or between sites. WSM reveals more information than just the information

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 1, April 2013

contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. Thus WSM pertains more information than just the information contain in a single document.

C. Web Usage Mining (WUM)

While content mining and structure mining utilize the real or primary data on the web, usage mining mines secondary data generated by the users' interaction with the web. Web usage data includes data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse-clicks and scrolls, and any other data generated by the interaction of users and the web. WUM [8] plays a key role in personalizing space, which is need of the hour.

III. LIMITATIONS OF EXISTING WEB MINING METHODS

As the amount of information on the web is increasing and changing rapidly without any control. As a result, the existing systems find difficulty in handling the newly emerged problems .Existing web mining is having some of the limitations so we move to soft web mining. Some of the problems [9] are discussed here.

(i) *Subjectivity, Imprecision, and Uncertainty*
(ii) *Deduction*
(iii) *Soft Decision*
(iv) *Clustering*
(v) *Dynamism, Scale, and Heterogeneity*
(vi) *Outlier*

IV. FUZZY LOGIC

Fuzzy logic [4,5] is a form of multi-valued logic derived from fuzzy set theory to deal with reasoning that is approximate rather than precise. In contrast with "crisp logic" where binary sets have binary logic, the fuzzy logic variables may have a membership value of not only 0 or 1 – that is, the degree of truth of a statement can range between 0 and 1. It is not constrained to the two truth values of classical propositional logic. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions.

The ability to model imprecise and qualitative knowledge and handle uncertainty are distinguished characteristics of fuzzy sets. Fuzzy logic is capable of addressing approximate or vague notions that are inherent in many information retrieval (IR) tasks. Fuzzy logic overcomes sharp boundary problems of many systems. An example of non fuzzy set and fuzzy set for age is given below in figure 1 and figure 2.

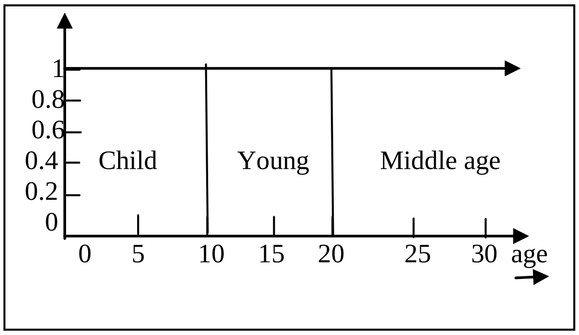

Fig. 1. Non Fuzzy Set

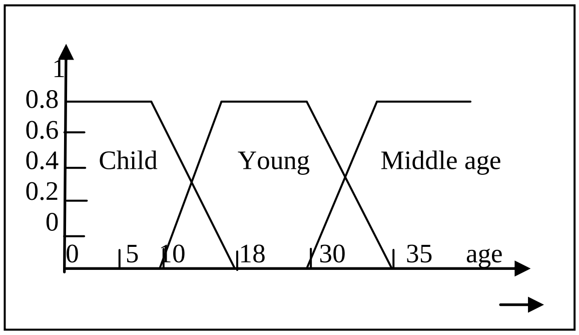

Fig. 2. Fuzzy Set

A. Importance of Fuzzy Clustering

The Information Retrieval system retrieves documents based on a given query. Both the documents and in most cases, the queries, are instances of natural language. Natural langue is often vague and uncertain. It is difficult to judge something that is vague and uncertain with deterministic crisp formulas and/or crisp logical rules. Fuzzy logic is based on the theory of fuzzy sets, a theory which relates to classes of objects with un-sharp boundaries in which membership is a matter of degree. Documents, queries and their characteristics could easily be viewed as fuzzy granular classes of objects with un-sharp boundaries and fuzzy memberships in many concept areas .Since the concept of fuzzy logic is quite intuitive, the fuzzy logic model provides a framework that is easy to understand for a common user of IR system.

In conventional clustering all objects either belongs to a category or not at all. But this is not always true because exception is there. With the use of fuzzy clustering, the documents can belong to more than one domain topic, represented by one group, with varying degrees of relevance. The relevance of documents with relation to groups can be represented by means of linguistic terms, which resembles in a more appropriate way the indication of importance given by human beings. For example, a document can belong "very much", or a "bit" to a particular group/topic or a topic can be "very important", or "minor" for the user's query. So fuzzy clustering is a better solution for handling such type of situations. In the next different fuzzy clustering techniques and their comparison with the conventional clustering has been discussed.

B. Hard C-Means (HCM) , Fuzzy C-Means (FCM) and Fuzzy C-Mediods (FCMedd)

In this section, HCM, FCM and FCMedd clustering techniques has been discussed so that they can be used to make clusters with unsharp boundaries.

In **HCM** [16] the function is defined as
Let C be the number of clusters, 1< C < N ,
B= { β_1 , β_2,…. β_C } be the prototypes of the clusters,
X = { x_1,x_2,…..x_c}be the set of N feature vectors, and U=[u_{ij}] be the fuzzy C-partition matrix. The objective function of the HCM algorithm is defined in equation (i)

$$J_H(B,X) = \sum_{i=1}^{C} \sum_{x_j=\beta_i}^{N} d_{ij}^2 \quad \text{(i)}$$

where d_{ij}^2 is the squared distance from point x_j to prototype β_i .The above equation can be reformulated as(iv)

$$J_{HCM} = \sum_{i=1}^{N} \min d_{ij}^2 = \sum_{i=1}^{N} (d_{min}^2)_j$$

(ii)

where $(d_{min}^2)_j$ is the squared distance from point x_j to nearest prototype.

In **FCM** [16] the objective function is defined as

$$J_H(B,X) = \sum_{i=1}^{C} \sum_{x_j=\beta_i}^{N} u_{ij}^m d_{ij}^2$$

(iii)

where

$$u_{ij} = \frac{d_{ij}^{2/(1-m)}}{\sum_{k=1}^{C} d_{kj}^{2/(1-m)}}$$

(iv)

and the fuzzifier $m \in (1,\infty)$. Substituting equation (iv) into (iii),we obtain

$$J_{FCM} = \sum_{j=1}^{N} (\sum_{i=1}^{C} d_{ij}^{2/(1-m)})^{1-m} = C \sum_{j=1}^{N} h_j^2$$

(v)

where h_j^2 is the harmonic mean of the distance d_{ij}^2 , i=1…..C,given by

$$h_j^2 = \frac{1}{C} \sum_{j=1}^{c} d_{ij}^{2/(1-m)})^{1-m}$$

Since C is a constant, it can be ignored in (v). Moreover, in our experiments, we will consider the m=2 case. Therefore, the objective function of the FCM can be written as

$$J_{FCM} = \sum_{j=1}^{N} h_j^2$$

(vi)

where $h_j^2 = \frac{1}{\sum_{i=1}^{C} \frac{1}{d_{ij}^2}}$

Thus, the FCM objective function tries to minimize the summation of harmonic mean distance h_j^2 of every feature vector to all clusters.

In **FCMedd** [22] the objective function is defined as:
Let X={x_i|i=1,…..,n} be a set of n objects. Each object may or may not represent by a feature vector. Let r(xi,xj) denote the dissimilarity between x_i and object x_j. Let V={v_1,v_2,…..v_c}, $V_i \epsilon X$ represent a subset of X with cardinality c, i.e.,V is a c-subset of X. Let X_C represent the set of all c-subsets V of X. The fuzzy mediods algorithm minimizes:

$$Jm(V;X) = \sum_{i=1}^{n} \sum_{i=1}^{c} u_{ij}^m r(x_j, v_i)$$

(vii)

Where the minimization is performed over all V in X_c. In (vii) u_{ij} represent the fuzzy or probabilistic membership of x_j in cluster i. The membership u_{ij} can be defined heuristically in many different ways. For example the above FCM membership model is given by:

$$U_{ij} = \frac{(\frac{1}{r(x_j,v_i)})^{1/(m-1)}}{\sum_{k=1}^{c} \quad (\frac{1}{r(x_j,v_k)})^{1/(m-1)}}$$

(viii)

Where m ϵ [1,∞] is the “Fuzzifier”. Another possibility is:

$$U_{ij} = \frac{\exp\{-\beta r(x_j,v_i)\}}{\sum_{k=1}^{c} \exp\{-\beta r(x_j,v_k)\}}$$

(ix)

Above equations generate a fuzzy partition of X in the sense that the sum of memberships of an object x_j across classes equal to 1. If we desire probabilistic memberships, we could use functions of the following type:

$$U_{ij} = [1 + \frac{r(x_j,v_k)}{\eta_i}]^{-1}$$

(x)

$$U_{ij} = \exp(-\frac{r(x_j,v_i)}{\eta_i})$$

(xi)

Since u_{ij} is a function of the dissimilarities $r(x_j, v_k)$, it can be eliminated from (vii). This is the reason j_m is shown as a function of *V* alone, When (vii) is minimized, the *V* corresponding to the solution generates a fuzzy or probabilistic partition via an equation such as (viii). However, (vii) cannot be minimized via the alternating optimization technique.

C. Analysis of HCM, FCM & FCMedd

In non fuzzy or hard C-means (HCM) clustering, data is divided into crisp clusters, where each data point belongs to exactly one cluster. No overlap between clusters is there. In HCM clusters can't be empty and can't contain all data points.

In this clustering technique partial membership is not allowed .HCM is used to classify data in a crisp sense. By this we mean that each data point will be assigned to one and only one data cluster. In this sense, these clusters are also called as partitions that are partitions of the data. In case of hard c mean each data element can be a member of one and only one cluster at a time.

FCM is an iterative algorithm. The aim of FCM is to find clusters centers (cancroids) that minimize a dissimilarity function. In Fuzzy clustering each member is associated some membership value, that indicate the strength of association between a data element and a particular cluster. FCM find clusters centers that minimize a dissimilarity function. FCM iteratively moves the cluster centers to the "right" location within a dataset.

Fuzzy set allows for degree of membership

A single point can have partial membership in more than one class.

There can be no empty classes and no class that contains no data points.

FCM iteratively moves the cluster centers to the "right" location within a dataset.

FCMedd is a well known algorithm that has objective function that is robust in nature. In other words, a single outlier object could lead to a very unintuitive clustering result. To overcome this FCMedd is being there. This algorithm is robust in nature because the performance is not effected by the presence outliers.

In measuring complexity

n number of data points
d number of dimensions
c number of clusters
i number of passes over entire dataset

Comparison between HCM,FCM and FCMedd is shown in the Table 1.

Table 1. Comparison between HCM, FCM and FCMedd

Comparison basis	**Hard C-Means**	**Fuzzy C-Means**	**Fuzzy C-Mediods**
Set type	Crisp set	Fuzzy set	Fuzzy Set
Speed	Faster	Slower	Much faster
Convergence	Slow	Fast & always converge	Very quickly
Belongingness of each data point to a cluster	Exactly one cluster	More than one cluster	More than one cluster with membership value
Overlapping	No overlapping is there	Overlapping is its advantages	Overlapping is possible
Clusters	Non Empty	Non empty	Non Empty
Sensitive	noise & outlier can be there	Less sensitive to noise & outlier	More sensitive to noise & outlires
Membership	Full	Partial	Partial
Computational time	Short	long	Long
Time Complexity	O(ncdi)	O(ndc^{2}i)	O(n^2)

V. CONCLUSION

Considering the immense potential of application of soft computing to web mining, this paper is timely and appropriate. In this paper, we have summarized the different types of web mining and its basic components, along with their current states of art. The limitations of the existing web mining methods/tools are explained. The relevance of soft computing and importance of fuzzy logic is already discussed. The possible future directions of using FL for some of these tasks are given in detail. Last, the use of Hard C-means(HCM), Fuzzy C-means(FCM) and Fuzzy C-Mediods (FCMedd) clustering are discussed in detail. Fuzzy clustering, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. Thus it is found that how these fuzzy clustering overcomes the disadvantages of the conventional methods that was used earlier.

References

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

[2] Christopher D. ManningPrabhakar Raghavan Hinrich Schütze" An Introduction to Information Retrieval"

[3] S. B¨utcher, C. Clarke, and G. Cormack."Information Retrieval: Implementing and Evaluating Search Engines". MIT Press, Cambridge, USA, 2010.

[4] M.Ganesh Introduction to FUZZY SETS & FUZZY LOGIC,PHI learning private limited

[5] J.Klir/Bo Yuan,Fuzzy sets and Fuzzy logic,Prentice Hall of India Private limited.

[6] Jiawei Han,Micheline Kamber,Data mining ,Morgan Kaufmann publishers

[7] K.Suresh "Improved FCM algorithm for Clustering on Web Usage Mining"

[8] Guandong Xu,Yanchun Zhang,Web mining & social networking

[9] Sankar K. Pal, Fellow, IEEE, Varun Talwar, Student Member, IEEE, and Pabitra Mitra, Student Member, IEEE "Web Mining in Soft Computing Framework:Relevance, State of the Art and Future Directions".

[10] Anjali B.Raut,G.R.Bamnote"Web Document Clustering".

[11] Dragos Arotariteia, Sushmita , Aalborg University Esbjerg, Niels Bohrs “Web mining: a survey in the fuzzy framework” Vej 8, 6700 Esbjerg, Denmark

[12] Xiuqi Li ,NSF/FAU “Web Document Classification Based on Fuzzy Association” Multimedia Laboratory, Florida Atlantic University,Boca Raton, FL 33431, USA.

[13] Menahem Friedman,Moti Schneider “A new approach for Fuzzy Clustering of Web Documents”.

[14] Tatiane.M.Nogueira,Helooisa A.Camargo “Fuzzy rules for document classification to improve Information Retrieval “

[15] Bernadette Bouchon-Meunier, Marcin Detyniecki, Marie-Jeanne Lesot, Christoph Marsala, and Maria Rifqi “Real-World Fuzzy Logic Applications in Data Mining and Information Retrieval”.

[16] Jongwoo Kim, Raghu Krishnapuram and Rajesh Dave”On Robustifying the C-Means Algorithm”,University of Missouri,Columbia.MO 65203.

[17] E.H. Ruspini. “A new approach to clustering. Information and control”, 22-32.[65] K-means clustering algorithm data mining tutorial started by KINGSLEYTAGBO at 12-14-2004

[18] Raghu Krishnapuram,Anupam Joshi,Liyu Yi,”A Fuzzy Relative of the K-Mediods algorithm with application to web document and snippest clustering” Colorado school of Mines.

[19] K.Suresh R.Madana Mohana A.RamaMohan Reddy “Improved FCM algorithm for Clustering on Web Usage Mining”.

[20] Fabio crastani, Gabriella Pasi “Soft Information retrieval based on Fuzzy set & Neural network”.

[21] Binu Thomas and Raju G,”A Novel Clustering Method for Outlier Detection in Data Mining”.

[22] T.Velmurugan and T.Santhanam,”A Comparative Analysis between K-Mediods And Fuzzy C-Means clustering algorithms for statistically distributed data points” ,Arumbakkam,Chennai-600106,India.

[23] Choochart Haruechaiyasak, Mei-Ling Shyu “Web Document Classification Based on Fuzzy Association “, FL 33124, USA Shu-Ching Chen.