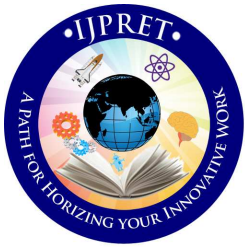
Research Article *ISSN: 2319-507X*
Anjali Raut, IJPRET, 2013; Volume 1(8): 507-520 *IJPRET*



INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

A FUZZY RECOMMENDER SYSTEM INTEGRATING WEB CONTENT AND WEB USAGE MINING

ANJALI B. RAUT, DR. G. R. BAMNOTE

1. **Department of Computer Science & Engg., HVPM's COET, Amravati, India.**
2. **Department of Computer Science & Engg., PRMITR, Badnera, Amravati, India.**

Accepted Date:

27/02/2013

Publish Date:

01/04/2013

Keywords

Web Mining,

World Wide Web,

Fuzzy logic (FL),

Web Content Mining,

Web Usage Mining.

Abstract

In today's era of e-commerce Recommender systems have become an important part. Business has come to realize the potential of these personalized and adaptive systems in order to increase sales and to retain their customers. Also web users have to rely on such systems to help them to in more efficiently and conveniently finding information of their interest from large information available on WWW. In this paper we propose a framework for recommender system which integrates fuzzy clustering of web pages and web usage mining approaches.

Corresponding Author

Ms. Anjali B. Raut

I. Introduction

OVER the last decade, we have witnessed an explosive growth in the information available on the World Wide Web (WWW). Today, web browsers provide easy access to many sources of text and multimedia data. More than billion pages are indexed by search engines, and finding the desired information is not an easy task. This large amount of resources creates the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "web mining." Web Mining is the use of Data Mining techniques to automatically discover and extract information from web. Web creates the new challenges of information retrieval as the amount of information on the web and number of users using web growing rapidly.

A recommender system is a web-based interactive software agent. It attempts to predict user references from user data and/or user access data for the purpose of facilitating and personalizing users' experience on-line by providing them with recommendation lists of suggested items. The recommended items could be products, such as books, movies, and music CDs, or on-line resources, such as web pages or on-line activities. Recommender systems have become extremely common in recent years. A few examples of such systems:

- When viewing a product on Amazon.com, the store will recommend additional items based on a matrix of what other shoppers bought along with the currently selected item.

- Netflix offers predictions of movies that a user might like to watch based on the user's previous ratings and watching habits (as compared to the behavior of other users), also taking into account the characteristics (such as the genre) of the film.

Recommender systems typically produce a list of recommendations in one of two ways - through collaborative or content-based filtering. Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviors, activities or preferences and predicting what users will like based on their similarity to other users. Content-based filtering methods are based on information about

and characteristics of the items that are going to be recommended. In other words, these algorithms try to recommend items that are similar to those that a user liked in the past (or is examining in the present). The objective of this paper is to provide an outline of web mining, its various classifications, Web Mining Process and propose a framework a for recommender system which integrates fuzzy clustering of web pages and web usage mining approaches.

II. Web Mining

World Wide Web is a major source of information. 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; there by 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.

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].

Web mining 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 data. This makes the

techniques to be used for a particular task in web mining widely varying. Some of the characteristics of web data are

1) Unlabeled
2) Distributed
3) Heterogeneous
4) Semistructured
5) Time varying
6) High dimensional.

Therefore, web mining basically deals with mining large and hyper-linked information base having the abovementioned characteristics. Thus, web mining, though considered to be a particular application of data mining, requires a separate field of research, mainly because of the abovementioned characteristics.

A. Web Mining Process

Web mining may be decomposed into the following subtasks:

1. *Information Retrieval/ Resource Discovery:* It is the process of retrieving the web resources. Resource discovery or IR deals with automatic retrieval of all relevant documents, while at the same time ensuring that the nonrelevant resources are fetched as few as possible. The IR process mainly includes document representation, indexing, and searching for documents.
2. *Information Extraction*: It is the transform process of the result of resource discovery and automatically extracting specific information from newly discovered Web resources. Once the documents have been retrieved the challenge is to automatically extract knowledge and other required information. Information extraction (IE) is the task of identifying specific fragments of a single document that constitute its core semantic content.
3. *Generalization:* It is process of uncovering general patterns at individual Web sites and across multiple sites [3]. In this phase, pattern recognition and machine learning techniques are usually used on the extracted information. Most of the machine learning systems, deployed on the web, learn more about the user's

interest than the web itself. A major obstacle when learning about the web is the labeling problem as data is rich on the web but it is unlabeled.

4. *Analysis:* Analysis is a data-driven problem which presumes that there is sufficient data available so that potentially useful information can be extracted and analyzed. Humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium. This is especially important for validation and or interpretation of the mined patterns which take place in this phase. Once the patterns have been discovered, analysts need appropriate tools. Some others use Online analytical processing (OLAP) techniques such as data cubes for the purpose of simplifying the analysis of usage statistics from sever access logs.

Based on the aforesaid four phases (Fig. 1), web mining can be viewed as the use of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services.

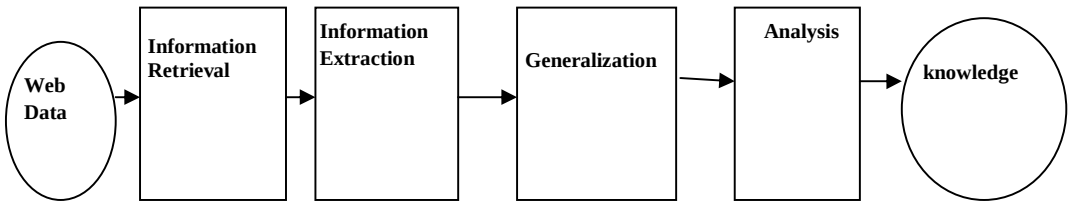


Figure 1. Web mining Process

B. Web Mining Taxonomy

Web Mining is use of Data Mining techniques to automatically discover and extract information from web documents and services [1].

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

- Web pages consist of text.
- Web pages are linked via hyperlinks
- 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][5][6]. Following Fig 2 shows the Web Mining taxonomy.

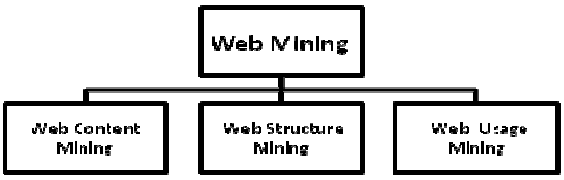


Figure 2. Web mining Taxonomy

C. 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.

D. 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. Web structure mining studies the web's hyperlink structure. It usually involves analysis of the in-links and out-links of a web page, and it has been used for search engine result ranking.

E. 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. Web usage mining focuses on analyzing search logs or other activity logs to find interesting patterns. One of the main applications of web usage mining is to learn user profiles.

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.

III. A Web Mining Framework for Fuzzy Recommender System

Figure 3 shows a Web Mining framework for fuzzy recommender system based on web content mining and web usage mining. The overall process is divided into two phase ,The online phase which consists of data preparation and web mining task and online phase which is real time recommendation engine.

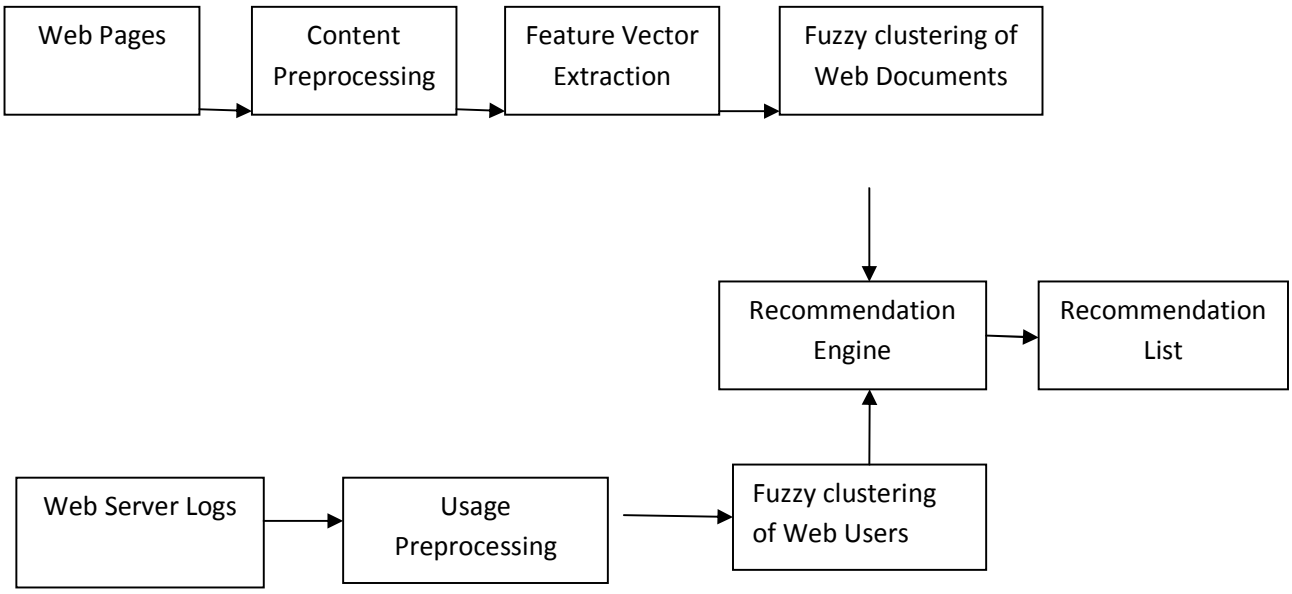


Figure 3. Web mining framework for fuzzy recommender system

IV. Clustering of Web Documents

Clustering analysis is widely used to establish object profiles based upon objects' variables. Objects can be customers, web documents, web users, or facilities (Chang, Hung, and Ho, 2007). Unlike classification, which analyzes class-labelled data objects, clustering analyzes data objects without consulting a known class label. The objects are clustered based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. This means that clusters of objects are created for the objects within a

cluster have high similarity in comparison to one another, but are very dissimilar to the objects of other clusters. There exist many clustering algorithms, which can be classified into several categories, including partitioning methods, hierarchical methods. A partitioning method classifies objects into several one-level clusters, where each object belongs to exactly one cluster, and each cluster has at least one object. A hierarchical method creates hierarchical decomposition of objects. Based on how the hierarchy is formed, hierarchical methods can be classified into agglomerative (bottom-up) approaches and divisive (top-down) approaches. All these traditional methods are hard clustering method. In hard clustering data is divided into distinct clusters, where each data element belongs to exactly one cluster. In **fuzzy clustering** (also referred to as **soft clustering**), data elements can belong to more than one cluster, and associated with each element is a set of membership levels.

A. Preprocessing of Web Pages

This preprocessing step comprises of different methods like Content extractor, stop word removal and stemming .

- In Content extractor phase only the primary content of web page is get identified.
- Cleaning HTML, XML or SGML tags from the Web pages,
- Eliminating all punctuations like comma, full stop, quotation mark, etc., only except the underscore in-between words,
- Eliminating all digital numbers,
- Changing all characters to lower case and
- Eliminating the stops words, which are very common words such as a, and the etc.

B. Web Page Representation Using Vector Space Model

The vector space model is based on linear algebra and treats documents and queries as vectors of numbers, containing values

corresponding to occurrence of words (called here **terms**) in respective documents Let t be size of the terms set, and n be size of the documents set. Then all documents D_i, i = 1..n may be represented as t-dimensional vectors:

$D_i = [a_{i1}, a_{i2}, ..., a_{it}]$

where coefficients a_{ik} represent the values of term k in document D_i [Salton 89]. Thus both documents and terms form a **term-document matrix .** A(n×t). Rows of this matrix represent documents, and columns – so called **term vectors**

	T_1	T_2	....	T_t
D_1	w_{11}	w_{21}	...	w_{t1}
D_2	w_{12}	w_{22}	...	w_{t2}
:	:	:		:
:	:	:		:
D_n	w_{1n}	w_{2n}	...	w_{tn}

Figure 4: Document representation in the vector space model

We use tfidf approach for term weighting. *Then tfidf (term frequency / inverted documentfrequency) of term j in document I is defined by:*

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Tfidf weighs the frequency of a term in a document with a factor that discounts its importance when it appears in almost all documents. Therefore terms that appear too rarely or too frequently are ranked lower than terms that hold the balance and, hence, are expected to be better able to contribute to clustering results.

Fuzzy Clustering using Fuzzy C Means

Clustering is one of the Data Mining techniques to improve the efficiency in information finding process. Many clustering algorithms have been developed and used One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) Algorithm. The FCM algorithm attempts to partition a finite collection of n elements $X = \{x_1, ..., x_n\}$ into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centres $C = \{c_1, ..., c_c\}$ and a partition matrix $W = w_{i,j} \in [0,1],\ i = 1, ..., n,\ j = 1, ..., c$, where each element w_{ij} tells the degree to which element x_i belongs to cluster c_j . Like the k-means algorithm, the FCM aims to

minimize an objective function. The standard function is:

$$w_k(x) = \frac{1}{\sum_j \left(\frac{d(\text{center}_k, x)}{d(\text{center}_j, x)}\right)^{2/(m-1)}}.$$

which differs from the k-means objective function by the addition of the membership values u_{ij} and the fuzzifier m. The fuzzifier m determines the level of cluster fuzziness. A large m results in smaller memberships w_{ij} and hence, fuzzier clusters. In the limit m = 1, the memberships w_{ij} converge to 0 or 1, which implies a crisp partitioning. In the absence of experimentation or domain knowledge, m is commonly set to 2.

V. Web Usage Mining

Web Usage Mining is the type of Web mining activity that involves the automatic discovery of user access patterns from one or more Web servers. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. In Web usage mining, the primary Web resource that is being mined is a record of the requests made by visitors to a Web site, most often collected in a Web server log. The content and structure of Web pages, and in particular those of one Web site, reflect the intentions of the authors and designers of the pages and the underlying information architecture. The actual behavior of the users of these resources may reveal additional structure.

The data object of web fuzzy clustering is the web source matrix which represents the data objects attributes of the given web data set, but the direct processing data object of web fuzzy clustering is web fuzzy similarity matrix or web fuzzy equivalence matrix. So we should abstract web source data firstly, get web data matrix representing web objects attributes, and then transform it into web fuzzy similarity matrix or web fuzzy equivalence matrix which are suitable for web fuzzy clustering. In the end, we use web fuzzy clustering method on web fuzzy similarity matrix or web fuzzy equivalence matrix and obtain clustering results. The Web Fuzzy Clustering processing Model (WFCM) is shown in Figure 5[15].

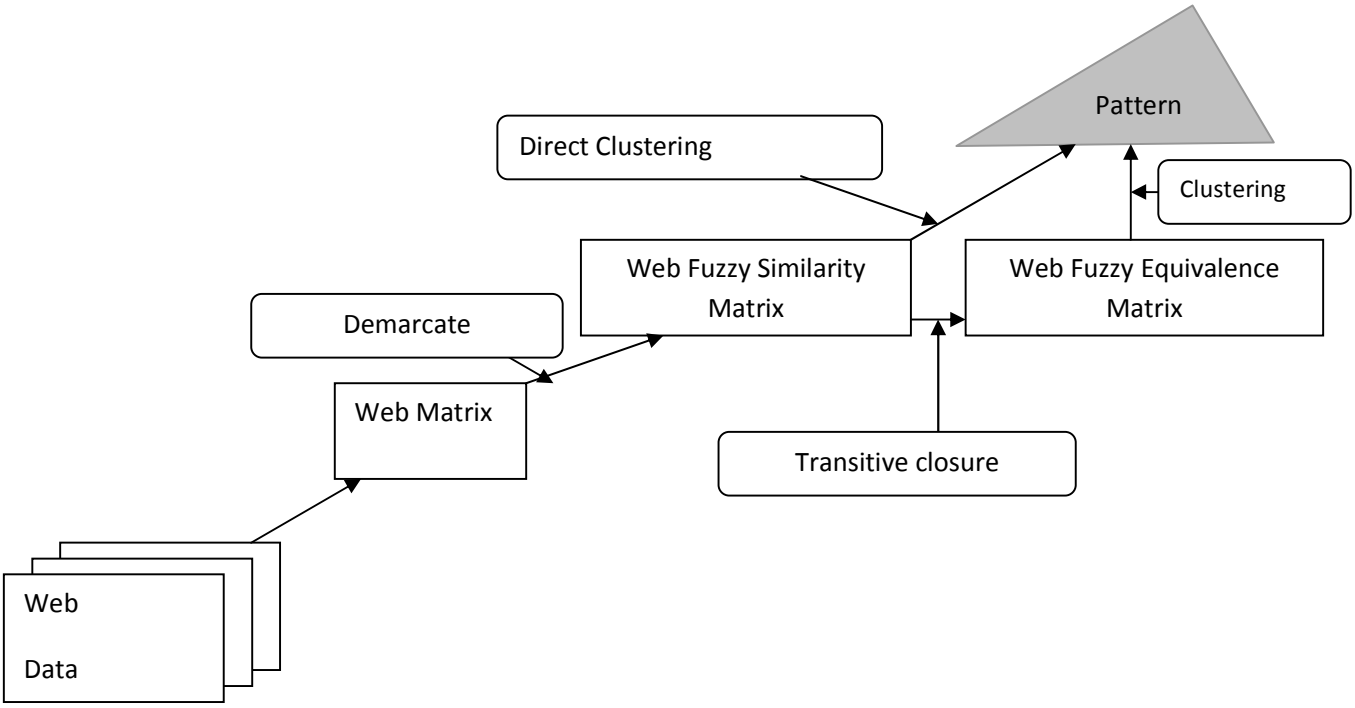


Figure 5: Web fuzzy clustering processing model

It is useful to combine Web usage mining with content and structure analysis in order to "make sense" of observed frequent paths and the pages on these paths. Following is a hybrid algorithm based input from fuzzy Web document clustering and Web Usage Mining. The proposed algorithm as follows.

Proposed Algorithm for Fuzzy Recommender System:

Input:

(1)Number of Clusters and Alpha Cut Value.

(2)Output from Fuzzy clustering of Web Documents i.e cluster of web pages.

(3) Output From Fuzzy clustering of Web users i.e. clusters of users.

Output:

Pr is the final recommend set of pages

Let Ci is current user and cluster set C={C1,C2,…Ck} consist of the members of same cluster Let {p1,p2…pk} are the pages visited by the users C

For (i=1; i<=k; i++)

{*// traverse the recommend set Pu=[p1,p2,…,pn] that come from Web Usage mining*

p=traverse (Pu);

//search the theme that p belongs to in the cluster sets

Topic=search1 (ClusterSets,p);

// K is the number of topic that P belongs to;

for (j=1; j<=k; j++)

{P1=search2(ClusterSets, topic);

//return the most N pages that similar to p from the theme;

P2= P2+P1;} P3= P3+ P2;}

//function Find finds other pages that similar to [p1, p2, …, pk] in same theme;

*//function Intersect look for the intersection between P and P3.*P'=Find(P3)

Pr=Intersect (P',P)

V. CONCLUSIONS

This paper proposes a Fuzzy Recommender System which uses web mining framework. It integrates the fuzzy clustering of web document and Web Usage mining for the discovery and analysis of Web navigational patterns based on fuzzy clustering of web documents and users. We believe that the successful integration of soft computing techniques with Web is likely to lead to the next generation of personalization tools and Recommendation System which are more intelligent and more useful for Web users.

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).

16. Klir & Yuan ,"Fuzzy Sets and Fuzzy Logic: Theory and Applications", Prentice Hall Publication.