An Optimal Clustering Approach In Web Usage Mining For Recommendation System

Tripti Saxena , Dr Pratima Gautam

Abstract: Web usage mining is used to identify and understand patterns that are taken from web data. Web usage mining is one of the applications of Data mining procedures to perceive and comprehend the requirements of electronic uses. In other words, the quick development of web-based business has incited the item to be over-burden, where the clients on the web are not ready to successfully choose the items they are presented to. To overcome these issues, web usage mining can be used to generate patterns. Web usage mining is an effective technique for extracting knowledge from unstructured data. In our research work, we are using HMM ranking and Expectation Minimization-Gaussian Mixture Model (EM-GMM) clustering for generating better patterns for the ease of web based services. The research work is performed on MATLAB simulation tool for generating rules for clusters first and second.

Keywords: Web Mining, Web usage mining, Clustering, HMM, Recommender System, Expectation Minimization-Gaussian Mixture Model (EM-GMM).

1. INTRODUCTION

Web usage mining (WUM) defines as web log mining as well. WUM mines the log data put away out in the web server. Colossal improvement in World Wide Web expands the multifaceted nature for clients to peruse it efficiently. To build the execution of sites better web architecture, web server exercises are moved to according to clients' interests. The ability to know the habits and interests of users supports in the operational approaches of enterprises.Web mining is the use of data mining processes to separate web data from web records, including web archives, hyperlinks between records, logging sites, and so on. Web mining (WM) can also be termed as the integration of the knowledge collected by conventional data mining methods and techniques with knowledge related to the web.Two unique methodologies were taken in at first characterizing Web mining. The initial one was a 'procedure driven view', which terms WM as a series of errands and 'data-driven view' was other one, in which Web mining is characterized as far as the kinds of Web data which was utilized in the mining procedure.

There are three common category of knowledge which knows how to be found by web mining:

- Web activity, server log & tracking the web browser activity.
- Web graphs, from connections to pages.
- For data found in web content, web pages and archives.

2. WEB USAGE MINING

With the increasing demand of internet more number of websites is being involved for getting required information and thus more usage of web-based data. The data which is stored in different format types in the web log file. This log record ought to be kept up as this data is in an unsorted way and it is done through preprocessing. WUM centers on finding valuable data. Web log record is consequently produced by the web server at whatever point client gets to the asset like site page of the site WUM is the way toward pulling back the helpful information from the server logs. It is the utilization of data mining procedures to find fascinating use designs from Web data so as to understand and better serve the prerequisites of the Web-based applications. Web usage data note down the identity of the user and their browsing behavior at a particular Web site. Usage data can be documented in the form of log files. A weblog is a record where the server takes the learning always client asks for websites from a specific Server. Log document might be put in 3 distinct areas for example web servers, web intermediary server, client's browser.

- web server log documents
- The Log document comes under Web-server focuses on customer's action, which gets the web-server via web-browser for a website.
- Web Proxy Server
- It's the middle server (mode of cooperation) that exists between the customer and Web server. With these lines, if the web server receives the customer's demand through the intermediary Server, next records for Log would be data of the proxy-Server & none customer's first. These web proxy servers keep individual Log record for the client's social event.
- Customer/User Browsers

These log documents can be created to remain in the customer's browser window. There are several software that a user can download in their browsing window. And, after all its say and the log records are available in the customer's browser window, the sections of Log document are only made by the web-server.

3. DESCRIPTION OF WEB USAGE MINING

Data mining process used by WUM to finding useful patterns from web usage data. Through log records and site related mining, it can be able to find user access patterns. Figure 3.1 displays the overall process of mining the Web usage with all essential steps. These are data collection, pre-processing, pattern discovery and pattern analysis.

- *Tripti Saxena is currently pursuing PhD in computer science and engineering in AISECT University, Bhopal (MP), PH-9753485817. E-mail:triptisaxena16@gmail.com*
- *Dr Pratima Gautam is currently Dean & Head in computer science & engineering in AISECT University, Bhopal (MP), PH-7354071699. E-mail: pratima_shkl@yahoo.com*

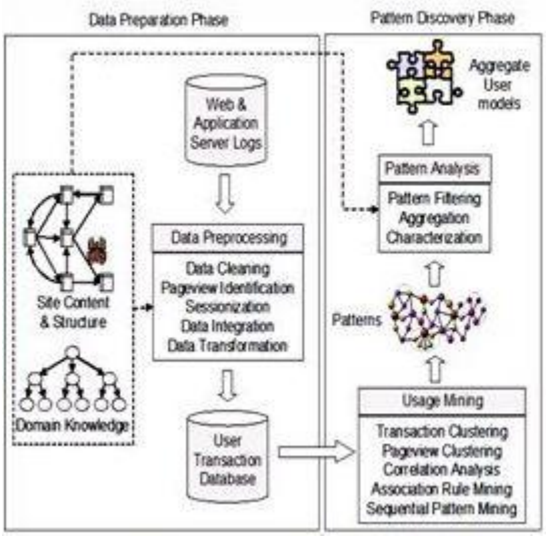


Figure 3*.1 Web Usage Mining steps*

Weblogs on web server are important data centers for WUM. When the user requests the resources of the web server, then each request is entered in the web log file on the web server. Consequently, clients browsing behavior is entered into the weblog document. Similarly, data is collected from website documents and operational database. The collected data in the weblog record is inadequate and is not suitable for direct mining. Pre-processing is important for transforming data into a suitable structure for pattern search. Pre-processing can give accurate, compact data for data mining, Includes data pre-processing, data cleaning, user identification, customer session recognition, path completion, and data integration [1]. The last step of the WUM process is the Pattern Analysis. Pattern discovery is the most important step of WUM process to extract the set of patterns and next choose the useful pattern and filter out uninteresting patterns by pattern analysis. By using some methods pattern analysis performed. These methods are data and knowledge query (like SQL), OLAP technology and usable analysis. After getting the patterns from queries then extracted pattern are arranged in cube then OLAP activities are carried out on these patterns. When patterns are extracted from pattern set there is need of representing the results, the methods of representation, for example, are used regularly for different properties to separate the coloring pattern or color to tilt the normal pattern or data. To filter the patterns (like specific page, content type page or hyperlinks) used content and structure information. The result of pattern analysis helps in e-commerce in the following ways:

- For fulfill the requirement of users according to user's browsing behavior from the sites, E-commerce web site pattern are improved.
- As of personal interest personalize the web sites and for the visitor change the web site specifically.
- To detecting the intrusion develop a security system which can be prohibited user access from online content.
- Clients can be maintained by understanding the client's basic requirements with providing customized products, satisfied the clients by the use of tracking behavior in e-commerce..
- Analyze the behavior patterns of consumers to effective estimation of advertise.

Web server activity includes server log files. Gives insights regarding document solicitations to a web-Server reaction to those solicitations, Weblogs are kept up by Web-Servers & hold data about clients getting to the webpage.
To know the user's behavior uses web log files. Weblog files have requests to address web servers and it can be recorded in the sequential request. Web Server logs are basic contented (ASCII) documents, which are autonomous of the server phase.
Web server logs are of following types:

- Access Log File: Information about all received requests & client's Server.
- In Access Log Records all the requests proceeds via Server.
- Error Log File: record of inner error. When there is an error, on the page the client is requesting the entry from the web server, which is entered in the Error Log. Mostly, Access & Error Logs are utilized, although agents and Referrer Logs can be enabled on the Server.
- Agent Log File: Knowledge regarding user's browser, browser edition.
- Referrer Log File: it is used to giving information of links and user is forwarded to the website.

Log files maintain various types of information in different web servers. Table 1 represents the some information about web log

IP address of User	IP address of User
Client's verification name	Username and password if he server necessitates client verification
The date-time stamp	The time spent on each website page when surfing by the site of the client
The HTTP request	The HTTP status code returned to the client
The Response status	Response status of demanded webpage
The size of requested recourses	Requested resource size
Reference URL	The asset gotten to by the client
Client's browser recognition	Browser wherever client sent a requests to the Web-Server
Visiting Path	The way the customer was taken whereas going to the site
Demand type	Strategy to be used for posted information is noted

4. RECOMMENDER SYSTEM

The Recommender System (RS) generally creates a list of one of two different traditions - in whole collaborative or content-based filtering. Content based algorithm RS [8] is the recommender framework that works with customers' profiles those are made towards the start. A profile contains knowledge regarding user and its taste. Taste of the user depends on user-rated items. In the recommended procedure, the motor analyzes things that are determined at the point along with the things fixed by the customer, which he does not rate and finds similarities. Customers who are strongly rated for the most part like this, the customer will be recommended.Recommender systems have most useable method which is Collaborative filtering Algorithm [8,6]. The design of CF is in discovering clients in a network that share thanks. On the off chance that two clients have the same or practically the same rated things in like manner, at that point they have comparable tastes[11]. Such clients fabricate a gathering or an alleged neighborhood. A client gets suggestions to those things that he/she hasn't evaluated previously, yet that was at that point emphatically appraised by clients in his/her neighborhood.Different methodologies of CF are:

- User-based methodology: This method was found in the late 1990s by the University of Minnesota's teacher Jonathan L. It was proposed by Heralkar. In the user-based methodology, clients do primary tasks. If some customers have the same taste then they are involved in a group. Depending on the assessment of things by different clients, an identical gathering is recommended to the customer, with who regularly share the inclination. If community rated an item positively, it would be suggested for the users [12,13].
- Item-Based methodology: In 2001, researchers of University of Minnesota projected this method [14]. Alluding to the way that the essence of clients stays consistent or alter somewhat comparative things fabricate neighborhoods dependent on thanks of clients. A while later the framework produces recommendations with things in the area that a client would incline toward [14] [23]
- Hybrid recommendation methodology: For better outcomes some recommender frameworks consolidate diverse procedures of community-oriented methodologies and content-based techniques.

We have proposed a new method of clustering namely Gausssian Mixture model (GMM) in this paper, which provides a significant improvement in the generation of no. of rules. Fig. 2 displays the role of recommendation system in web usage mining.The remainder of this paper is arranged as follows. Segment V provides preprocessing and point-by-point information of data mining methods that are accessible for use. In Segment VI, the review of the whole proposed framework is talked about pursued by the usage subtleties and result assessment in segment VII. Segment VIII portrays the end alongside future work.

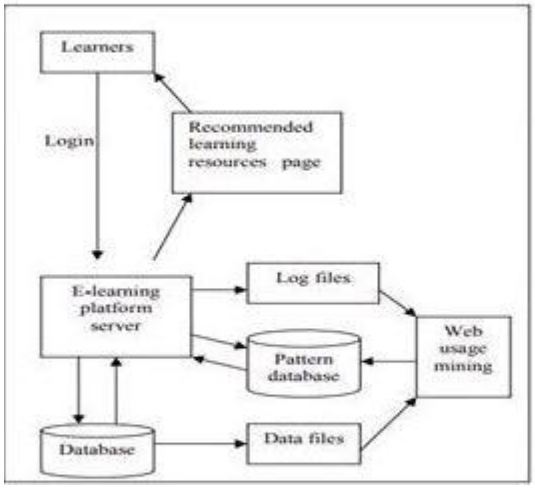


***Figure 2** Web Usage Mining based Recommender System*

5. LITERATURE REVIEW

Cooley et al introduced web usage mining firstly, and it focused on users' preferences on the Internet to prediction and learning [7]. The whole procedure of WUM is commonly partitioned into two essential assignments: data planning & pattern discovery [2]. The essential data hold by some servers such as Web servers, proxy servers and web customers, which is required for WUM. [9] It has been estimated web log data preprocessing takes 80% time of the data mining. Preprocessing work can be done in two ways: In the primary system, weblogs are mapped to relational databases; subsequently suitable mining calculations are adjusted for more analysis [10]. The other strategy uses an extraordinary preprocessing way to change the log notification to fit clear mining algorithms. Incidentally, the proposed technique in this study uses the first approach to pre-processing data. The data preparation work creates a server session file, where each session contains a sequence of different types of requests made by a user during the same visit [2]. For access web log data different preprocessing function are used. In [11], portrayal web browsing designs of data planning techniques. Various strategies are discussed about finding designs used to be specific Apriori [12], Naesve Bayesian [13], and agglomerative clustering [14] and they are given actual form. The pattern discovery functions include the findings of association rules, sequential patterns, user classification etc. [15]. The use of removed utility designs from web data can be related to broad scope, for example, optimized personalization, framework enhancement, webpage modification, business intelligence search, usage characterization, and so on [16]. On the basis of sales data, customer related data and web log data recommended product in this research. [17]. Different types of recommended techniques have been identified and established in many years. Collaborative filtering is one of the method of recommendation. In [18], author used NP-Miner Algorithm to store received web data on navigation pattern. In view of this data ongoing suggestions are given to online clients. The investigation exhibits that this computation beneficially plays out an online special recommendation relentlessly. In [19] a customized proposal framework for shopping from online is depicted. Web usage data, association rules, item logical classification, and decision tree induction are used for better recommendations on this structure. in this analysis tries to give a suitable recommendation for all visitors of commercial sites those having some characteristics to enrolled or unregistered. This exploration work goings-on for enhancing the nature of suggestion to not registered clients hence these clients are furnished by customized inclination. This strategy is beneficial to gather old users as well as the one-time user who visit the site the first time.Web utilization data mining could connect new clients, keep up current clients, track clients who are leaving a site, etc [20]. Use knowledge could be evacuated to extend web-Server profitability via prefetching with saving methods [21]. In light of a couple look at completed in the field of Web Mining, this could widely arrange it into 3 regions: Web Content Mining, Web Structure Mining, & Web Usage Mining. Web Content Mining is the way toward extricating learning from web archives, for example, content and sight and sound.Extracting the information of web structure and references of hyperlink termed as web structure mining. WUM is the process of misuse of information from Assistant Information [2]. With helpful information we mean that the browser Logs, client profile, Web-Server Logs, enlistment information, customer session or exchanges, behavior, customer inquiries & other type knowledge which is effect of web connection.In [8] A fuzzy rule-based clustering algorithm uses a managed sequence method so that the unsupervised bunch can handle the probe. It naturally attempts to investigate potential groups in data designs and identify them with some interpretable fuzzy rules. Together with these fuzzy rules, classification of data designs can highlight the actual boundaries of the clusters.

6. PROPOSED WORK

The Gaussian mixture model (GMM) based clustering are proposed by using enhance the probability function with EM algorithm [5], [6], [7], [8] which gives better results in compare of Fuzzy C-Means (FCM). This proposed method is more suitable to find the arbitrary ellipsoidal shapes clusters with arbitrary number of data points.In GMM-EM clustering use the existing HMM ranking results as a input parameter. [5-8] accommodating GMM-EM to increasing probability function with respect to parameters (components and mixing coefficient of covariances and involved average). The steps of EM clustering have been presented further. µkα) By considering the parameters for the current GMM-EM stage, the means μ_k covariance matrices$\sum_k$ and combination coefficients π_k(where k=1,....K) as a result of the previous period of k-means clustering calculate and assess the initial value of log probability..β) E step. Assess the obligations utilizing the present values of parameter

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n|\mu_k \sum_k)}{\sum_{i=1}^{k} \pi_l N(x_n|\mu_l \sum_l)} (1)$$

Ƴ) M step. By using current responsibilities again calculate the parameters [5-8]

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \quad (2)$$

$$\sum_l = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \quad (3)$$

$$\pi_k = \frac{N_k}{N} \quad (4)$$

Where,

$$\sum_{n=1}^{N} \gamma(z_{nk}) = N_k \quad (5)$$

δ) Assess the log probability

$$\ln p(X| \pi, \mu, \sum) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sum_k) \quad (6)$$

ε) Parameters convergence or log probability tests. On the off chance that the convergence criterion isn't fulfilled, come back to step b.

HIDDEN MARKOV MODEL

HMM depends on Markov series. The Markov model is used to defining the random variable probability and the state's sequence, which is one of the few who can go against prices from some set. These sets can be words, or labels, or images speaking to anything, similar to the climate. A Markov chain makes a solid presumption that on the off chance that we need to anticipate the future in the sequence, the only thing that is important is the present state. The states before the current state have no impact on the future except via the current state. It's as though to anticipate tomorrow's climate you could inspect the present climate yet you weren't permitted to see yesterday's climate.All the more formally, take a series of state factors q1,q2,...,qi . A Markov model encapsulates the Markov suspicion on the probabilities of this arrangement: that Markov presumption while foreseeing the future, the past doesn't make a difference, only the present. Markov supposition:

$$P(qi = a|q1...qi-1) = P(qi = a|qi-1) \quad (7)$$

A Markov chain is valuable when we have to process a probability for an arrangement of noticeable events. Much of the time, be that as it may, the events we are keen on are concealed: we don't watch them specifically. For instance we don't regularly watch grammatical feature labels in content. Or maybe, we see words and should induce the labels from the word succession. We call hidden tags because they are not observed. A hidden Markov model (HMM) enables us to discuss both events seen (such as we meet in information) and hidden incidents (such as grammatical form labels) which we consider to be causal factors in our potential model. A HMM is determined by the accompanying parts:

S = $s_1 s_2$ ...s_M is set of M states

M = m11m12 ...mn1 ...mnnis a transition probability matrix M, all mkl speaking to the moving probability from state P k to state l, subjected to $\sum_{l=1}^{n}$ mkl = 1 ∀ k

UO = uo1uo2 ...uosis U observations chain, every one drained from a dictionary D = d1, d2,..., di

OP = opi(uot) is observation probabilities chain, additionally called as outflow probability, all communicating the likelihood of a perception ot being produced from a state i

π = π1,π2,...,πNis a starting probability dispersion in excess of states. The probability of πi that Markov series will start in the state i. some of the states from j have πj = 0. These states cannot be considered as start states. Additionally, $\sum_{i=1}^{n}$ πi = 1.

Proposed Algorithm:

```
Step:1 Initializethe process
Step:2 Apply preprocessing
Step:3 AssessSession and customer identification
Step:4 Apply calculated Knowledge Base and Pattern mining
Step:5 Now rules will be generated using Apriori Algorithm
         Apply Apriori Algorithm on selected data
         JOINING PHASE: $K_i$ is generate by join the $F_{i-1}$ with itself
         PRUNING PHASE: some (i-1)-item set that e not repeated most of the time, is not considerable as frequent i-item set subset
         $K_i$: Candidate item set with i size
         $F_i$: Frequent item set with i size
         $F_1$ = {Frequent Item};
         For ( i = 1; $F_i$ != null; i++) do begin
         $C_{i+1}$ = Generated Candidates from $F_i$ ;
         For every T transactions in DB do
         Add the all candidates count in $C_{i+1}$ which are enclosed in T
         $F_{i+1}$ = candidates in $K_{i+1}$ with min_supp
         End for
         Returns $\cup_i F_{i;}$
Step:6 Generate PID for all existing product
Step:7 Calculate PID counts for visited URLs and Purchased
       Products
Step:8 If Counts >Threshold
Step:9 Evaluate Min. Supp Value (MSV) and Min. Conf. Value
       (MCV)
Step:10 Now apply EMGMM over the data
Step:11 Calculate PID counts for visited URLs and Purchased
       Products
Step:12 Form two clusters of the generated URLs and
       Products
Step:13 Rule formation performed over the clusters
Step:14 End the process
```

The above pseudocode shows the basic steps of the entire process which highlights the sequence of the techniques applied over the system. The threshold value taken in step 8 is visit =30 and purchase = 03.

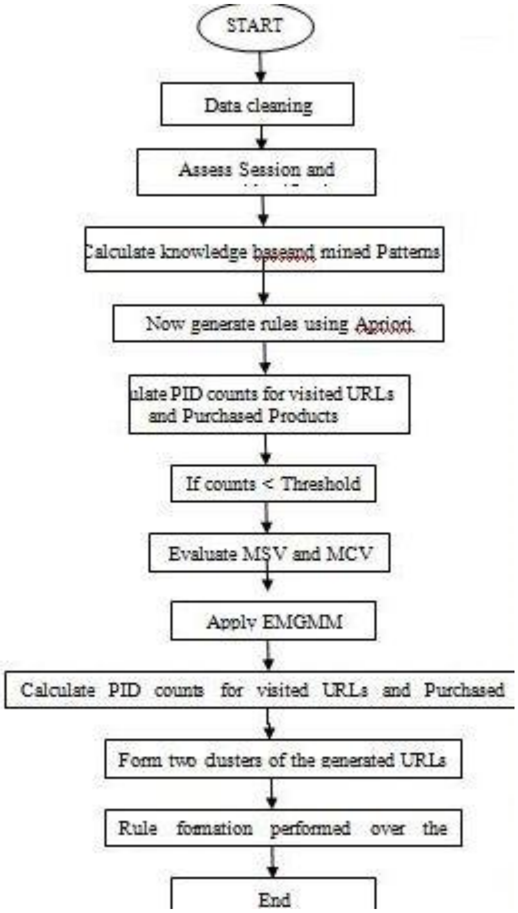

START
Data cleaning
Assess Session and
Calculate knowledge basesand mined Patterns
Now generate rules using Apriori
ulate PID counts for visited URLs and Purchased Products
If counts < Threshold
Evaluate MSV and MCV
Apply EMGMM
Calculate PID counts for visited URLs and Purchased
Form two clusters of the generated URLs
Rule formation performed over the
End

***Figure 3** Data Flow Diagram of EMGMM*

MSV = 0.12 and MCV = 0.2.

7. RESULT ANALYSIS

MATLAB uses a wide assortment of potential along with signal and image handling, correspondence, control configurations, test sizing, financial modeling and analysis, computational science and parallel preparation. The current PC framework has standard anti-focus and then-throughput-based stimulants, for example, large enlisting power in the form of normal processing units. The MATLAB programs are definite and expressly communicate information level equality as language gives some unusual state administrators who work specifically on groups. Traditionally, MATLAB is used as a programming language for writing different types of simulations. In areas such as control engineering, image processing and communication, it is widely used for simulation and designing the system. These programs are usually long lasting and developers make significant efforts in trying to reduce their running time.The rules produced in excess of the cluster once mining association rules from this sort of non-value-based data, we may discover hundreds or thousands of rules compared to explicit feature values. We in this manner present a clustered association rule as a rule that is shaped by consolidating comparative, \adjacent" association rules to structure a few general rules. Figure 4 and figure 5 shows the rules generated without EMGMM.

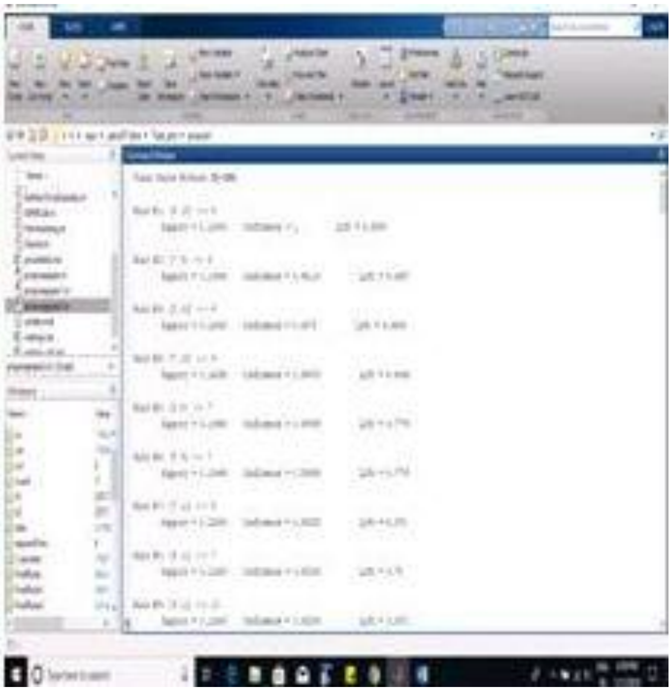
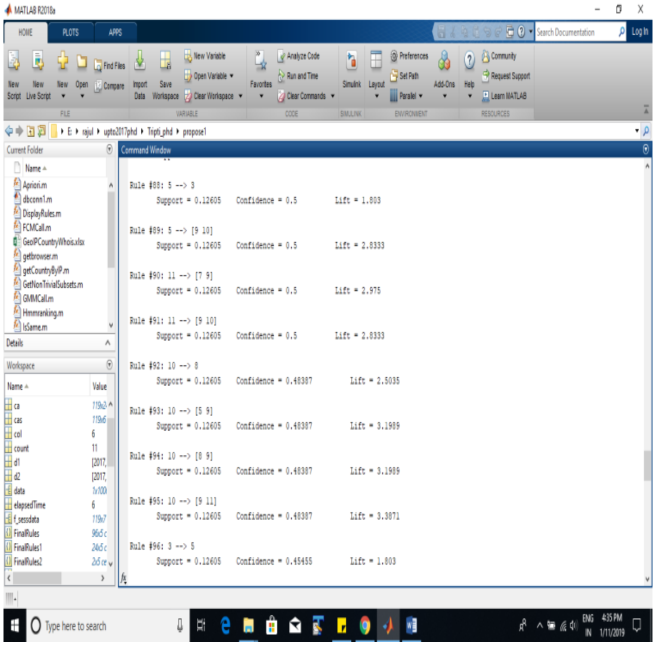


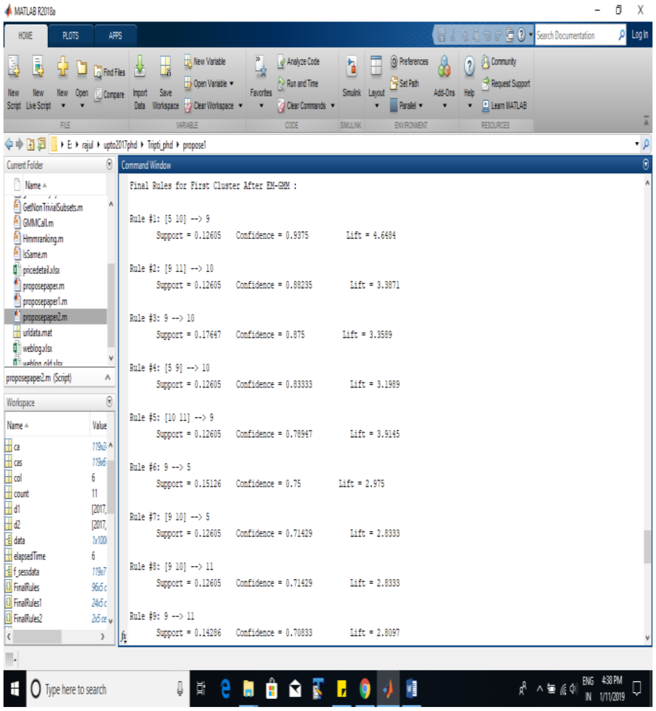
***Figure 4** visualizing the rules generated without EM-GMM*



```
Rule #88: 5 --> 3
      Support = 0.12605    Confidence = 0.5         Lift = 1.803

Rule #89: 5 --> [9 10]
      Support = 0.12605    Confidence = 0.5         Lift = 2.8333

Rule #90: 11 --> [7 9]
      Support = 0.12605    Confidence = 0.5         Lift = 2.975

Rule #91: 11 --> [9 10]
      Support = 0.12605    Confidence = 0.5         Lift = 2.8333

Rule #92: 10 --> 8
      Support = 0.12605    Confidence = 0.48387         Lift = 2.5035

Rule #93: 10 --> [5 9]
      Support = 0.12605    Confidence = 0.48387         Lift = 3.1989

Rule #94: 10 --> [8 9]
      Support = 0.12605    Confidence = 0.48387         Lift = 3.1989

Rule #95: 10 --> [9 11]
      Support = 0.12605    Confidence = 0.48387         Lift = 3.3871

Rule #96: 3 --> 5
      Support = 0.12605    Confidence = 0.45455         Lift = 1.803
```

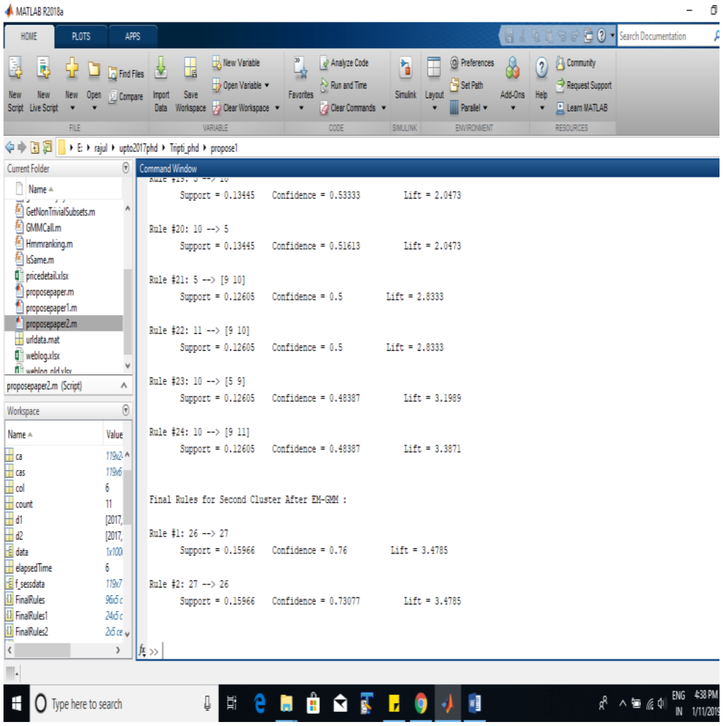
***Figure 5** visualizes continued phase of rules generated without EM-GMM*

Clustering technique will help to cluster the scattered data under a specific category which will help to work on rule generation more efficiently and will help to extract more useful and relevant rules. Figure 6 shows the generation of rules for first clsuter after applying EM-GMM clustering technique. Figure 7 shows the generation of rules for second cluster after applying EM-GMM



```
Final Rules for First Cluster After EM-GMM :


Rule #1: [5 10] --> 9
      Support = 0.12605    Confidence = 0.9375         Lift = 4.6484


Rule #2: [9 11] --> 10
      Support = 0.12605    Confidence = 0.88235         Lift = 3.3871


Rule #3: 9 --> 10
      Support = 0.17647    Confidence = 0.875         Lift = 3.3589


Rule #4: [5 9] --> 10
      Support = 0.12605    Confidence = 0.83333         Lift = 3.1989


Rule #5: [10 11] --> 9
      Support = 0.12605    Confidence = 0.78947         Lift = 3.9145


Rule #6: 9 --> 5
      Support = 0.15126    Confidence = 0.75         Lift = 2.975


Rule #7: [9 10] --> 5
      Support = 0.12605    Confidence = 0.71429         Lift = 2.8333


Rule #8: [9 10] --> 11
      Support = 0.12605    Confidence = 0.71429         Lift = 2.8333


Rule #9: 9 --> 11
      Support = 0.14286    Confidence = 0.70833         Lift = 2.8097
```

***Figure 6** shows the rules generated for first cluster after EMGMM*



```
Rule #19: 5 --> 10
      Support = 0.13445    Confidence = 0.53333         Lift = 2.0473


Rule #20: 10 --> 5
      Support = 0.13445    Confidence = 0.51613         Lift = 2.0473


Rule #21: 5 --> [9 10]
      Support = 0.12605    Confidence = 0.5         Lift = 2.8333


Rule #22: 11 --> [9 10]
      Support = 0.12605    Confidence = 0.5         Lift = 2.8333


Rule #23: 10 --> [5 9]
      Support = 0.12605    Confidence = 0.48387         Lift = 3.1989


Rule #24: 10 --> [9 11]
      Support = 0.12605    Confidence = 0.48387         Lift = 3.3871



Final Rules for Second Cluster After EM-GMM :


Rule #1: 26 --> 27
      Support = 0.15966    Confidence = 0.76         Lift = 3.4785


Rule #2: 27 --> 26
      Support = 0.15966    Confidence = 0.73077         Lift = 3.4785


>>
```

***Figure 7** shows the rules generated for second class after EM-GMM*

Final Rules Without EM-GMM :

Rule #1: [8 10] --> 9

Supp = 0.12605 Conf = 1 Lift = 4.9583

Rule #2: [7 8] --> 9

Supp = 0.13445 Conf = 0.941 Lift = 4.666

Rule #3: [5 10] --> 9

Supp = 0.126 Conf = 0.937 Lift = 4.648

Rule #4: [7 10] --> 9

Supp = 0.142 Conf = 0.894 Lift = 4.436

Rule #5: [5 9] --> 7

Supp = 0.134 Conf = 0.888 Lift = 3.777

Rule #6: [8 9] --> 7

Supp = 0.134 Conf = 0.888 Lift = 3.777

Rule #7: [7 11] --> 9

Supp = 0.126 Conf = 0.88235 Lift = 4.375

Rule #8: [9 11] --> 7

Supp = 0.126 Conf = 0.88235 Lift = 3.75

Rule #9: [9 11] --> 10

Supp = 0.126 Conf = 0.88235 Lift = 3.387

Rule #10: 9 --> 10

Supp = 0.17647 Conf = 0.875 Lift = 3.3589

Rule #11: [7 9] --> 10

Supp = 0.14286 Conf = 0.85 Lift = 3.2629

Rule #12: [5 7] --> 9

Supp = 0.13445 Conf = 0.84211 Lift = 4.1754

Rule #13: 9 --> 7

Supp = 0.16807 Conf = 0.83333 Lift = 3.5417

Rule #14: [5 9] --> 10

Supp = 0.12605 Conf = 0.83333 Lift = 3.1989

Rule #15: [8 9] --> 10

Supp = 0.126 Conf = 0.833 Lift = 3.198

Rule #16: 15 --> 14

Supp = 0.15966 Conf = 0.82609 Lift = 3.9322

Rule #17: 22 --> 17

Supp = 0.15126 Conf = 0.81818 Lift = 4.0568

Rule #18: [9 10] --> 7

Supp = 0.14286 Conf = 0.80952 Lift = 3.4405

Rule #19: [7 9] --> 5

Supp = 0.13445 Conf = 0.8 Lift = 3.1733

Rule #20: [7 9] --> 8

Supp = 0.13445 Conf = 0.8 Lift = 4.1391

Rule #21: 40 --> 41

Supp = 0.126 Conf = 0.789 Lift = 4.4737

Rule #22: [10 11] --> 9

Supp = 0.12605 Conf = 0.78947 Lift = 3.9145

Rule #23: 8 --> 9

Supp = 0.15126 Conf = 0.78261 Lift = 3.8804

Rule #24: 16 --> 17

Supp = 0.15126 Conf = 0.78261 Lift = 3.8804

Rule #25: 4 --> 3

Supp = 0.142 Conf = 0.772 Lift = 2.786

Rule #26: 14 --> 15

Supp = 0.159 Conf = 0.76 Lift = 3.93

Rule #27: 26 --> 27

Supp = 0.159 Conf = 0.76 Lift = 3.478

Rule #28: 9 --> 5

Supp = 0.151 Conf = 0.75 Lift = 2.97

Rule #29: 9 --> 8

Supp = 0.151 Conf = 0.75 Lift = 3.880

Rule #30: 17 --> 16

Supp = 0.151 Conf = 0.75 Lift = 3.88

Rule #31: 17 --> 22

Supp = 0.151 Conf = 0.75 Lift = 4.05

Rule #32: [7 9] --> 11

Supp = 0.126 Conf = 0.75 Lift = 2.97

Rule #33: 8 --> 7

Supp = 0.142 Conf = 0.739 Lift = 3.141

Rule #34: 27 --> 26

Supp = 0.159 Conf = 0.730 Lift = 3.478

Rule #35: 6 --> 7

Supp = 0.126 Conf = 0.714 Lift = 3.035

Rule #36: 7 --> 9

Supp = 0.168 Conf = 0.714 Lift = 3.54

Rule #37: 41 --> 40

Supp = 0.126 Conf = 0.714 Lift = 4.47

Rule #38: [9 10] --> 5

Supp = 0.126 Conf = 0.714 Lift = 2.83

Rule #39: [9 10] --> 8

Supp = 0.126 Conf = 0.714 Lift = 3.69

Rule #40: [9 10] --> 11

sup = .126 conf. = .714 LIFT = 2.80

Rule #41: 2 --> 5

sup = .142 conf. = .708 LIFT = 2.83

Rule #42: 9 --> 11

sup = .142 conf. = .708 LIFT = 2.80

Rule #43: 13 --> 17

Supp = 0.142 Conf = 0.708 Lift = 3.51

Rule #44: 17 --> 13

Supp = 0.142 Conf = 0.708 Lift = 3.512

Rule #45: 9 --> [7 10]

Supp = 0.142 Conf = 0.708 Lift = 4.43

Rule #46: 8 --> 5

Supp = 0.134 Conf = 0.695 Lift = 2.75

Rule #47: 8 --> [7 9]

Supp = 0.134 Conf = 0.695 Lift = 4.13

Rule #48: 22 --> 13

Supp = 0.126 Conf = 0.681 Lift = 3.380

Rule #49: 22 --> 16

Supp = 0.126 Conf = 0.681 lift = 3.52

Rule #50: 7 --> 5

sup = 0.159 conf. = .678 LIFT = 2.69

Rule #51: 7 --> 10

sup = .159 conf. = .678 LIFT = 2.60

Rule #52: 10 --> 9

Supp = .176 Conf = 0.677 Lift = 3.35

Rule #53: 9 --> [5 7]

Supp = 0.134 Conf = 0.666 Lift = 4.175

Rule #54: 9 --> [7 8]

Supp = 0.134 Conf = 0.666 Lift = 4.666

Rule #55: 8 --> 10

Supp = 0.126 Conf = 0.652 Lift = 2.50

Rule #56: 16 --> 22

Supp = 0.126 Conf = 0.652 Lift = 3.52

Rule #57: 8 --> [9 10]

Supp = 0.126 Conf = 0.652 Lift = 3.69

Rule #58: 5 --> 7

sup =.159 conf. = .633 LIFT = 3.69

Rule #59: 11 --> 10

sup = .159 conf. = .633 LIFT = 2.43

Rule #60: 2 --> 11

up = .126 conf. = .625 LIFT = 4.479

Rule #61: 13 --> 22

Supp = 0.126 Conf = 0.625 Lift = 3.38

Rule #62: 9 --> [5 10]

Supp = 0.126 Conf = 0.625 Lift = 4.648

Rule #63: 9 --> [7 11]

Supp = 0.126 Conf = 0.625 Lift = 4.375

Rule #64: 9 --> [8 10]

Supp = 0.126 Conf = 0.625 Lift = 4.958

Rule #65: 9 --> [10 11]

Supp = 0.126 Conf = 0.625 Lift = 3.914

Rule #66: 10 --> 7

sup = .159 conf. = .612 LIFT = 3.604

Rule #67: 10 --> 11

sup = 0.159 conf. = .61 LIFT = 4.43

Rule #68: 7 --> 8

Supp = 0.142 Conf = 0.607 Lift = 3.14

Rule #69: 7 --> 11

Supp = 0.142 Conf = 0.607 Lift = 2.40

Rule #70: 7 --> [9 10]

Supp = 0.142 Conf = 0.607 Lift = 3.44

Rule #71: 5 --> 9

Supp = 0.151 Conf = 0.6 Lift = 2.97

Rule #72: 7 --> [5 9]

Supp = 0.134 Conf = 0.571 Lift = 3.77

Rule #73: 7 --> [8 9]

Supp = 0.134 Conf = 0.571 Lift = 3.77

Rule #74: 5 --> 2

sup = .142 conf. = .566 LIFT = 2.80

Rule #75: 5 --> 11

sup = .142 conf. = .566 LIFT = 2.24

Rule #76: 11 --> 5

sup = .142 conf. = .566 LIFT = 3.247

Rule #77: 11 --> 7

sup = .142 conf. = .566 LIFT = 3.40

Rule #78: 11 --> 9

sup = .142 conf. = .566 LIFT = 4.809

Rule #79: 10 --> [7 9]

Supp = 0.142 Conf = 0.548 Lift = 3.262

Rule #80: 7 --> 6

Supp = 0.126 Conf = 0.535 Lift = 3.035

Rule #81: 7 --> [9 11]

Supp = 0.126 Conf = 0.535 Lift = 3.75

Rule #82: 5 --> 8

sup = .134 conf. = .533 LIFT = 3.75

Rule #83: 5 --> 10

sup = .134 conf. = .533 LIFT = 3.047

Rule #84: 5 --> [7 9]

Supp = 0.134 Conf = 0.533 lift = 3.173

Rule #85: 10 --> 5

sup = .134 conf. = .516 LIFT = 3.04

Rule #86: 3 --> 4

supp = .142 conf. = .515 LIFT = 4.78

Rule #87: 11 --> 2

sup = .126 conf. = .501LIFT = 1.479

Rule #88: 5 -->3

sup = .12605 conf. = .521 LIFT = 1.803

Rule #89: 5--> [9 10]

Supp = 0.12605Conf = 0.5lift = 2.8333

Rule #90: 11 --> [7, 9]

Supp = 0.12605Conf = 0.5 Lift = 2.975

Rule #91: 11 -->[9.10]

Supp = 0.12605Conf = 0.5 Lift = 2.8333

Rule #92: 10 -->8

Supp = 0.12605Conf = 0.48387 Lift = 2.5035

Rule #93: 10-->[5 9]

Supp = 0.12605Conf = 0.48387 Lift = 3.1989

Rule #94: 10 -->[8 9]

Supp = 0.12605Conf = 0.48387 Lift = 3.1989

Rule #95: 10 -->[9 11]

Supp = 0.12605Conf = 0.48387 Lift = 3.3871

Rule #96: 3 -->5

Supp = 0.12605Conf = 0.45455 Lift = 1.803

Final Rules for First Cluster After EM-GMM :

Rule #1: [5 10] --> 9

Supp = 0.12605 Conf = 0.9375 lift = 4.6484

Rule #2: [9 11] -->10

Supp = 0.12605 Conf = 0.88235 lift = 3.3871

Rule #3: 9 --> 10

Supp = 0.17647 Conf = 0.875 lift = 3.3589

Rule #4: [5 9] --> 10

Supp = 0.12605 Conf = 0.83333 lift = 3.1989

Rule #5: [10 11] --> 9

Supp = 0.12605 Conf = 0.78947 lift = 3.9145

Rule #6: 9 -->5

sup= .15126 conf = .75 LIFT = 1.975

Rule #7: [9 10] -->5

sup = .12605 conf. = .71429 LIFT = 1.8333

Rule #8: [9 10] -->11

sup = .12605 conf. = .71429 LIFT = 3.8333

Rule #9: 9 --> 11

sup = .14286 conf. = .70833 LIFT = 4.8097

Rule #10: 10 --> 9

Supp = 0.17647 Conf = 0.67742 lift = 3.3589

Rule #11: 11 -->10

Supp = 0.15966 Conf = 0.63333 lift = 2.4312

Rule #12: 9 -->[5 10]

Supp = 0.12605 Conf = 0.625lift = 4.6484

Rule #13: 9 --> 8

Supp = 0.12605 Conf = 0.625 lift = 3.9145

Rule #14: [7 9] --> 11

sup = .15966 conf. = .6129 LIFT = 1.4312

Rule #15: 5 -->9

sup = .15126 conf. = .601 LIFT = 3.975

Rule #16: 5 -->11

sup = .14286 conf. = .56667 LIFT = 3.2478

Rule #17: 11 -->5

sup = .14286 conf.= .56667 LIFT = 3.2478

Rule #18: 11 -->9

sup = .14286 conf. =0.56667 LIFT = 4.8097

Rule #19: 5 --> 10

sup = .13445 conf. =0.53333 LIFT = 4.0473

Rule #20: 8 --> 5

sup = .13445 conf. = .51613 LIFT = 1.0475

Rule #21: 5 --> [9 10]

sup = .12605 conf. = .502 LIFT = 1.83333

Rule #22: 11 -->[9 10]

sup = .12605 conf. = .512 LIFT = 3.8333

Rule #23: 10 --> [5 9]

Supp = 0.12605 Conf = 0.48387 lift = 3.1989

Rule #24: 10 --> [7 8]

Supp = 0.12605 Conf = 0.48387 lift = 3.3871

Final Rules for Second Cluster After EM - GMM :

Rules #1: 26 --> 27

Supp = 0.15966 Conf = 0.76 Lift = 3.4875

Rule #2: 27 -->26

Supp = 0.15966 Conf = 0.73077 Lift = 3.4785

8. CONCLUSION & FUTURE SCOPE

as with web-based applications growth, there is a need for understanding web usage data better especially in e-commerce, it is an important interest in analyzing web usage data and applying knowledge for providing the best use for users. Still, WUM arises some hard scientific questions, which should be answered before developing strong equipment. In this research work, we have provided a brief scenario of web usage mining in recommendation system using clustering techniques. This task will develop a structure or a idea which introduces the generation of rules over the clusters. It is a new concept of generating the rules over the clusters. In future we will try to apply some other technique to develop new and more efficient rules and also we will try to perform our research on some other advance simulation tool.

REFERENCES

[1] E. Kim, W. Kim, Y. Lee. Purchase propensity prediction of EC customer by combining multiple classifier based on GA. International Conference on Electronic Commerce 2000 : 274~280.

[2] J. B. Schafer, J. A. Konstan, J. Riedl. E-commerce recommendation applications. Data Mining and Knowledge Discovery,2001(5):115~153.

[3] S. W. Changchien, T. Lu. Mining association rules procedure to support on-line recommendation by customers and products fragmentation. Expert Systems with Applications, 2001(20): 325~335.

[4] B. Sarwar, G. Karypis, J. Konstan, J. Riedl. Analysis of recommendation algorithms for e-commerce. Proceedings of ACM Ecommerce 2000 Conference, 2001:158~167.

[5] S. Yuan, W. Chang. Mixed-initiative synthesized learning approach for Web-based CRM. Expert Systems with Applications, 2001(20):187~200.

[6] R. Kohavi, F. Provost. Applications of data mining to electronic commerce, Data Mining and Knowledge Discovery, 2001(5):1~2.

[7] J. g. Liu, h. h. Huang. Web Ming for Electronic Business Application, Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies, Chengdu, China, 2003:872~876.

[8] J. Srivastava, R. Cooley, M. Deshpande, P. N. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations,2000, 1(2):1~12

[9] E. Cohen, B. Krishnamurthy, J. Rexford. Improving end-to-end performance of the web using server volumes and proxy filters. In Proc. ACM SIGCOMM, 1998:241~253.

[10] M. Eirinaki, M. Vazirgiannis. Web mining for web personalization. ACM Transactions on Internet Technology(TOIT), 2003(1): 1~27.

[11] P. Pirolli, J. Pitkow, R. Rao. Silk from a sow's ear: Extracting usable structures from the Web. In: Proc. 1996 Conference on Human Factors in Computing Systems (CHI-96), Vancouver, British Columbia, Canada, 1996.

[12] L. Catledge, J. Pitkow. Characterizing browsing behaviors on the World Wide Web, Computer Networks and ISDN Systems ,1995,27(6):1065~1073.

[13] M. S. Chen, J. S. Park, P. S. Yu. Data mining for path traversal patterns in a web environment. In Proceedings of the 16th International Conference on Distributed Computing Systems, 1996:385~392.

[14] B. Mobasher. Web Usage Mining and Personalization. Practical Handbook of Internet Computing,2004:14~22

[15] R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago, Chile, Sept 1994.

[16] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery: An overview. In Fayyad U., Piatetsky-Shapiro G., Smyth P.,Uthurusamy R, editors, Advances in Konwledge Discovery and Data Mining, AAAI/MIT Press, 1996:1~34

[17] J. E. Pitkow, K. B. Krishna. Webviz: A tool for world-wide web access log analysis. In First International WWW Conference, 1994.

[18] K. Hammond, R. Burke, C. Martin, s. Lytinen. Faq-Finder: A case-Based approach to knowledge navigation. In Working Notes of the AAAI Spring Sym posium: Information Gathering from Heterogeneous, ternational Distributed Environments. AAAI Press, 1995.

[19] T. Joachims, D. Freitag, T. Mitchell. WebWatcher: A Tour Guide for the World Wide Web, Proceedings of IJCAI97, Nagoya, Japan, August 1997:770~775.

[20] D.S.W. Ngu, X. Wu. Sitehelper: A localized agent that helps incremental exploration of the world wide web. In 6th International World Wide Web Conference, Santa Clara, CA, 1997. 710

[21] H. Lieberman. Letizia: An agent that assists web browsing. In Proc. of the 1995 International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995.