Characterizing Web Users Based on Their Required Criteria

Ming-Yi Shih
Department of Computer Science and Information Engineering
National Changhua University of Education
Changhua City, Taiwan, ROC
myshih@cc.ncue.edu.tw

Syun-Sian Huang
Department of Computer Science and Information Engineering
National Changhua University of Education
Changhua City, Taiwan, ROC
rexchem1990@gmail.com

***Abstract*—In order to run a successful website, it is significantly crucial for website owners to understand users' intentions and desires. By capturing these information, they can provide better service and enhance marketing strategy to achieve this goal. Web usage mining (WUM) is an application that can help people to explore the useful patterns of users' browsing usages. Traditionally, it discovers knowledge from Web log data. However in some websites, they offer a service that users can select or enter some required criteria from fields, and these information will be saved online. These criteria show the intentions or desires of a certain object required for this user. Interested persons can enter queries or browse categories to find these posted cases. In this paper, clustering method is applied to group similar users based on these collected required criteria in a website. When dataset is huge, it is difficult to find the characteristics of individual group. Thus association rule mining is applied to each cluster. The generated rules can be inferred to identify the interests and characteristics of users in each group. Finally, marketing decision can be made especially for each group's users.**

Keywords-Web usage mining; required criteria; clustering; pattern discovery

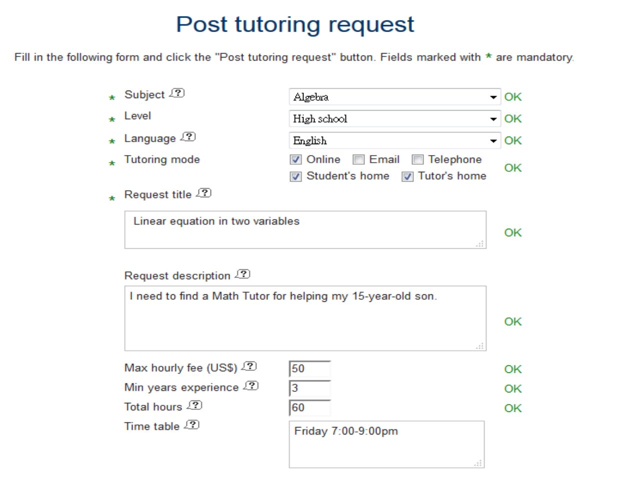
1. INTRODUCTION

As Web applications grow tremendously and versatilely in recent years, surfing Web has become an essential activity of daily life for many people. More and more people enjoy the provided information and services online. There is no doubt that Web is changing the way we live. Consequently, the business transactions carried out over the Web are significantly increased. Since most organizations do not ignore this trend, doing business online and running a successful website have become important goals in their business activities.

The data about the activities of users surfing a Web site are important sources for Web site administrators or webmasters. The knowledge of users' behaviors can be collected by analyzing these data. The goals are to improve the website's performance, to provide better services, and make effective marketing strategies. These are crucial tasks for running a successful website.

Web usage mining (WUM) is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications [27]. It has been intensively explored in recent years [2-5]. However traditional WUM researches focus on Web server logs. A typical Web server logs record the information about user's IP address, the access date and time, the URL of the requested page, the protocol, the return code, and the size of the page if it is a successful request. The browsing behaviors of Web users can be identified by processing Web server logs. Additionally, Y. H. Tao, T. P. Hong, and Y. M. Su [6] install software in client side to collect users' browsing activities, for example saving a file, clicking back on browser... etc., as an additional source for mining. Proxy side logs that collected from internet service providers can also be processed like Web server logs to find the behavior of navigations.

In this paper, a WUM application based on Web users' required criteria is proposed. When people surf online to find jobs, housing, tutoring … and so on, they can input search queries to look for objects on their own initiative. However lots of websites offer a service that users can select or fill value in provided fields as preferences and save these criteria online waiting for interested persons to find their case. For example, a user want to post a case to look for a tutors in a tutoring website [28]. He/she needs to enter required criteria, such as purpose, subject, the requirement for tutors (including teaching experience, hourly rate) …and so on. The parts of this Web page is shown as Figure 1.



Post tutoring request

Fill in the following form and click the "Post tutoring request" button. Fields marked with * are mandatory.

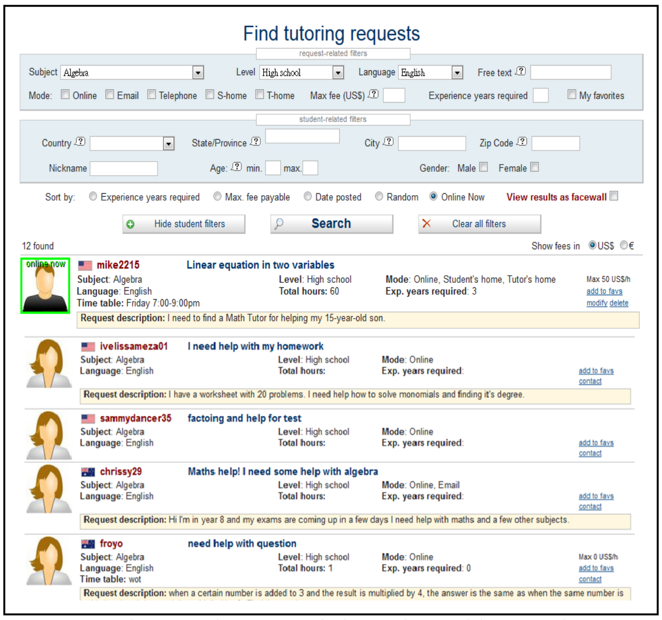
* Subject	Algebra	OK
* Level	High school	OK
* Language	English	OK
* Tutoring mode	☑ Online ☐ Email ☐ Telephone ☑ Student's home ☑ Tutor's home	OK
* Request title	Linear equation in two variables	OK
Request description	I need to find a Math Tutor for helping my 15-year-old son.	OK
Max hourly fee (US$)	50	OK
Min years experience	3	OK
Total hours	60	OK
Time table	Friday 7:00-9:00pm	

Figure 1: parts of Web page for entering required criteria

QSHINE 2015, August 19-20, Taipei, Taiwan

DOI 10.4108/eai.19-8-2015.2260878

These required criteria will be saved online. The interested tutors can find this posted case by searching or browsing this website. Parts of the searing results for above case are presented in figure 2.


Find tutoring requests

request-related filters

Subject: Algebra | Level: High school | Language: English | Free text:

Mode: □ Online □ Email □ Telephone □ S-home □ T-home | Max fee (US$): | Experience years required: | □ My favorites

student-related filters

Country: | State/Province: | City: | Zip Code:

Nickname: | Age: min. max. | Gender: Male □ Female □

Sort by: ○ Experience years required ○ Max. fee payable ○ Date posted ○ Random ● Online Now | View results as facewall □

12 found | Show fees in ● US$ ○ €

online now
mike2215 — Linear equation in two variables
Subject: Algebra | Level: High school | Mode: Online, Student's home, Tutor's home | Max 50 US$/h
Language: English | Total hours: 60 | Exp. years required: 3
Time table: Friday 7:00-9:00pm
Request description: I need to find a Math Tutor for helping my 15-year-old son.

ivelissameza01 — I need help with my homework
Subject: Algebra | Level: High school | Mode: Online
Language: English | Total hours: | Exp. years required:
Request description: I have a worksheet with 20 problems. I need help how to solve monomials and finding it's degree.

sammydancer35 — factoing and help for test
Subject: Algebra | Level: High school | Mode: Online
Language: English | Total hours: | Exp. years required:

chrissy29 — Maths help! I need some help with algebra
Subject: Algebra | Level: High school | Mode: Online, Email
Language: English | Total hours: | Exp. years required:
Request description: Hi I'm in year 8 and my exams are coming up in a few days I need help with maths and a few other subjects.

froyo — need help with question
Subject: Algebra | Level: High school | Mode: Online | Max 0 US$/h
Language: English | Total hours: 1 | Exp. years required: 0
Time table: wot
Request description: when a certain number is added to 3 and the result is multiplied by 4, the answer is the same as when the same number is

Figure2: the parts of view of searching results

These criteria present the behaviors of Web users about the intentions or desires of certain objects. Potentially useful knowledge of Web users' patterns is hidden inside these criteria. Since the number of Web users and the variety of their behavior, the discovery of every single user's access patterns is not feasible. WUM applications frequently apply clustering algorithms to group users with similar properties to conclude group of users' behaviors. The characteristic of each group can be achieved by analyzing these common properties. Nevertheless, the analysis of clustering results is still hard to find the hidden patterns sometimes. In this work, association rules mining is adopted to each group. The generated rules show the associations among items in criteria in each group. The characteristics of each group can be referred by examining these association rules of individual group.

This paper is organized as follows: In Section 2, related work and background are discussed. Section 3 outlines the proposed methods for finding characteristics of Web users based on their criteria. Experimental observations are presented in Section 4. Section 5 concludes this work.

2. RELATED WORK AND BACKGROUND

2.1 Clustering

Clustering is an important field of data mining, and has been widely used to divide the data set into several groups such that objects have a high degree of similarity to each other in the same group and have a high degree of dissimilarity to the ones in different groups [7]. Thus, clustering techniques has been applied to WUM applications to explore the behaviors of user groups.

In [8], a page hierarchy generated from URLs is applied to generalized user sessions that are extracted from Web log. BIRCH clustering algorithm [9] is used to clustering these generalized data. Since the dimension of data set is reduced, this method can yield fast and effective clustering results of Web users.

Y. Liu., and X. Huang [10] employ a mixture of Markov models to group users based on their historical access sequences. The sequential relationships inherent in user navigation histories can be captured by applying this model. Then, personalized recommendations to assist the user's navigation can be made by mined knowledge.

G. T. Raju, and M. V. Sudhamani, [11] propose a neural network based clustering algorithm and apply it to Web log files of NASA Web site to find users' access patterns.

E. Sadikov et al., [12] combines information from document-clicks and from user session as query refinements, and presents user search behavior as graph. Therefore patterns about user search behavior can be mined by performing multiple random walks on a Markov graph. The results can be used to query suggestion

Y. Zhang, and G. Xu [13] measures web similarity by finding the linking-relevant page groups via linkage structure analysis for clustering. This proposed algorithm combined with the latent semantic analysis to identify user access patterns.

J. Zhang et al., [14] improved the fuzzy clustering algorithm and applied it to Internet banking data to find to an interesting target groups.

K-means clustering algorithm [15] is one of the most popularly used clustering algorithms in Web usage mining application [16-18]. Therefore, the k-means algorithm is employed in this work.

2.2 Association rules mining

The goal of association rules mining is to identify sets of data attributes or items that are statistically related. An association rule usually is expressed as the form of A -> B, where A and B are two disjoint itemsets. For example, we may find the rule: "80% of customers who buy bread and egg will also buy milk" in the supermarket's transaction data sets. In this example, bread, egg and milk are attributes or items in transactions. In WUM applications, association rules mining are usually used to discover co-occurrence relationship among accessed Web pages.

B. Mobasher et al., [19] propose effective and scalable techniques for Web personalization based on association rule discovery from usage data.

M. Dimitrijević, and Z. Bošnjak [20] think that Web usage data differs from the market basket data. When two pages have a link between them, they may be accessed together. The support and confident of Web pages are higher than market basket data. Large number of candidate sets are generated during association mining. A schemes for pruning are proposed to reduce the size of candidate sets.

R. Suguna, and D. Sharmila [21] mine constraint association patterns from clustered user group, and assign a weight to be associated with each page in a transaction. These patterns will reflect interest of each page based on time of stay and quality rating. These generated rules can be used to assist users for browsing.

V. Nebot and R. Berlanga [22] present a method to derive appropriate transactions from semantic instance data repositories expressed in RDF/(S) and OWL by capturing the implicit schema-level knowledge encoded in the ontology. Association rules can be mined from these derived transactions.

Apriori algorithm [23] is one of the most famous algorithm for finding association rules. This algorithm has been intensively applied for Web usage mining [24-26]. Thus this algorithm is applied in this research to discover association rules.

3. CHARACTERIZING WEB USERS BASED ON THEIR REQUIRED CRITERIA

In order to find the characteristics of Web users based on their required criteria by applying clustering technique and association rule mining, an approach shown on figure 3 is proposed in this work.

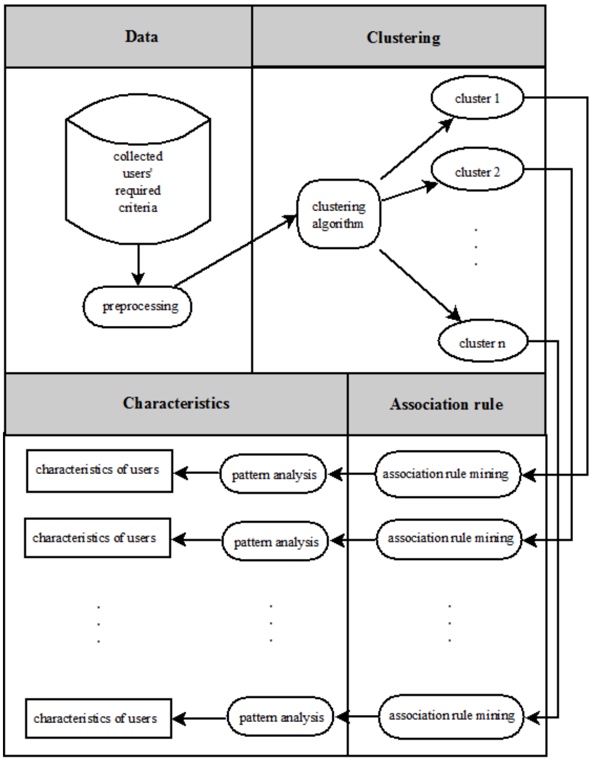


Figure 3: the framework of proposed approach

3-1 Data and data preprocessing

[29] is a famous website to find a tutor or tutoring jobs in Taiwan. Data used in this work are collected from this website during April to September 2008. The required criteria provided by students or their parents to find a tutor are focused. An interested tutor can find an interesting student according to these criteria. Since some fields of provided criteria (e.g. name, time of availability) are considered useless for mining users' properties, they will be ignored. Moreover, some value of fields are merged for simplification. The hourly rate that is the only numeric feature is transferred to categorical value, because original Apriori association rules mining algorithm only accepts this kind of data. The fields of collected criteria and their associative value are preprocessed as follows:

Students: preschooler, primary school, junior high school, senior high school, college/university, and adult.

Subjects: art/sport, daycare, Math/Physics/Chemistry, English, other academic subjects, third foreign languages, and technology/professional subjects.

Required teaching experience: none, one year, two years, three years, four years, and five years and above.

Hourly rate: low, medium, high.

3-2 Clustering and association rule mining

K-means clustering algorithm is a distance based method. Every object that needs to be input into this algorithm is usually viewed as a point in a multidimensional space. It is represented as vector space model, ($x_1, x_2, ... x_n$), where x_i is the value of i-th selected attribute. For example, (preschooler, daycare, none, low) may represents one collected data to be input into k-means in this case to find to which group it belongs.

Based on the experimental observations, the number of cluster is set to seven. Because this setting of cluster number show the properties of every cluster clearly. The association rules mining is applied to each cluster to identify characteristics of Web user. Nevertheless some of generated rules are defined as redundant, they were removed. For example, the followings are two generated rules in a cluster:

1. { Student = junior high school, subject= English, required teaching experience = none } -> { hourly rate = low }
2. { Student = junior high school, subject = English } -> { hourly rate = low }

Rule 2 is considered as a redundant rule, because rule 1 already provide a more precise information. Therefore, rule 2 will not appear in the final rule set. The rules that are defined as uninteresting were also pruned. Once the association rules are generated, the characteristics of each group can be inferred. The results are shown on next section.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

The rules found in each cluster and the observations about these rules are listed below.

Cluster 1:

1. { student = adult, subject = English, required teaching experience = none } -> { hourly rate = low }
2. { student=preschooler, subject = daycare required teaching experience=none }->{ hourly rate = low }
3. { subject = art/sport } -> {required teaching experience = none, hourly rate = low }
4. { subject = third foreign language, required teaching experience = none } -> { hourly rate = low }

People in this group want to learn English, art/sport, third foreign language, or seek daycare for their children with a limited budget and do not require teaching experience. It seems the hourly rate is the major concern for them, and the quality of service can be sacrificed.

Cluster 2:

1. { student = adult, subject = English, required teaching experience = none } -> { hourly rate = medium }
2. { student = adult, subject = art/sport } -> { hourly rate = medium }
3. { student = adult, subject = third foreign language } -> { hourly rate = medium }

These rules indicate that there are some adults need to develop or improve their skill in foreign language or in art/sport. They believe they can enhance their careers or lives in a variety of ways by learning these subjects. Therefore they are willing to pay a higher price to strengthen or broaden their expertise.

Cluster 3:

1. { student = adult, subject = English, required teaching experience = one year } -> { hourly rate = low }
2. { required teaching experience = two years } -> { hourly rate = low }
3. { student = preschooler, required teaching experience = one year } -> { hourly rate = low }
4. { subject = third foreign language, required teaching experience = one year} -> { hourly rate = low }
5. { student = senior high school, required teaching experience = one year } -> { hourly rate = low }

Although persons in this group are will to pay the low hourly rate, they require one-two years of teaching experience. It seems they do not want to scarify much quality of service. Most of them are adult seeking for language tutors, and parent looking for a daycare services or for a tutor for their senior high school children.

Cluster 4:

1. { student = adult, required teaching experience = one year, hourly rate = medium } -> { subject = English }
2. { student = adult, required teaching experience = two years, hourly rate = medium } -> { subject = English }
3. { student = adult, required teaching experience = three years} => { subject = English }

There are some adults need to learn English with higher budgets, and they think an experienced tutor may achieve better results during learning.

Cluster 5:

1. { subject = Math/Physics/Chemistry, student = senior high school } -> { hourly rate = medium }
2. { subject = English, student = senior high school } -> { hourly rate = medium }
3. { student = primary school, subject = English } -> { hourly rate = medium }
4. { student = primary school, subject = art/sport } -> { hourly rate = medium }
5. { student = junior high school, subject = English } -> { hourly rate = medium }
6. { student = junior high school, subject = Math/Physics/Chemistry } -> { hourly rate = medium }

The hourly rate is not the major concern for the parents in this group. They offer a higher tuition to find a tutor to help their children for schoolwork or art/sport.

Cluster 6:

1. { subject = Math/Physics/Chemistry, student = junior high school, required teaching experience = one year } -> { hourly rate = low }
2. { subject = Math/Physics/Chemistry, student = junior high school, required teaching experience = two years } -> { hourly rate = low }
3. { subject = English, student = junior high school, required teaching experience = one year } -> { hourly rate = low }
4. { subject = English, student = junior high school, required teaching experience = two years } -> { hourly rate = low }

This group of parents can be defined as wise users, because there are plenty of university students looking for tutoring jobs. They can assist junior high school students to strengthen schoolwork. Thus these parents think they can provide low hourly rate to find an experienced tutor without difficulty.

Cluster 7:

1. { student = college/university } -> { subject = technology/professional subjects }
2. { student = college/university, hourly rate = medium } -> { subject = English }

This group are university or college students need to enhance their expertise or English skill that are import for their future career or study.

Once the characteristics of each group is defined, some marketing strategies can be made based on above observations. The first idea that comes to mind is to sell certain items or services to the people that will most likely buy them. For example, an advertisement of an expensive computer aided English learning software designed for adult can be sent to the people in group 2 and 4. An English supplementary material for junior high school students can be promoted to the people in group 5 and 6. Moreover, some other target marketing strategies can be obtained in this research.

1. The persons in group 2, 4 and 5 are willing to pay a higher tuition, thus the probability for them to buy an expensive product or service is higher than other groups.
2. Because group 1, 3, and 6 contain the people trying to find a tutor with limited budget, selling expensive items to them may get little responses.

3. Group 3 and 6 consist of the persons that try to find an experienced tutor with limited budge. It seems they do not want to scarify much quality. Thus, discounted and necessary products or service may get interesting to them.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a method for characterizing Web users based on their required criteria. The Web users are clustered into seven groups, and the patterns of each group can be identified by applying association rules mining efficiently. We also advocate that the mined patterns from these users' entered criteria are fruitful properties for target marketing.

In this study, we try to find out Web users' behaviors based only on single data source, however it is feasible to combine Web log data, users' client side activities or users' registration information with these proposed data. The patterns of users about their Web activities can be captured versatilely. It will be the extension of this work.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] P. Nithya and P. Sumathi, "A Survey on Web Usage Mining: Theory and Applications," International Journal of Computer Technology and Applications, vol. 3 (4), pp. 1625-1629, 2012.

[3] N. H. Panchal and O. Kale, "A Survey on Web Usage Mining," International Journal of Computer Trends and Technology, vol. 17 (4), pp. 177-181, 2014.

[4] A. K. Singh, D. Sharma, and A. Pathak, "Web Usage Mining: A Concise Survey on Tools and Applications," International Journal of Computer Applications, vol. 74 (1), pp. 1-7, 2013

[5] T. Tabassum, and S Saxena, "A Survey on Web Usage Mining Techniques," International Journal of Engineering Research & Technology, vol. 2 (10), pp.3824-3829, 2013.

[6] Y. H. Tao, T. P. Hong, and Y. M. Su, "Web usage mining with intentional browsing data," Expert Systems with Applications, vol. 34 pp. 1893–1904, 2008.

[7] J. Han, and K. Kamber, "Data mining: Concept and Techniques," San Francisco: Morgan Kaufman Publisher. 2001.

[8] Y. Fu, K. Sandhu, and M. Shih., "A generalization based approach to clustering of web usage sessions," In International WEBKDD Workshop, San Diego, CA, 1999

[9] T. Zhang, R. Ramakrishman, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," Proc1996 ACM-SIGMOD Int. Conf. Management of Data, 1996 June. Montreal, Canada. 1996.

[10] Y. Liu., and X. Huang, "Personalized Recommendation with Adaptive Mixture of Markov Models," Journal of the American Society for Information Science and Technology, vol 58, (12), pp.1851-1870, 2007.

[11] G. T. Raju, and M. V. Sudhamani, "A novel approach for extraction of cluster patterns from Web Usage Data and its performance analysis", International Conference on Emerging Trends in Electrical and Computer Technology, Nagercoil, India, 2011

[12] E. Sadikov, J. Madhavan, L. Wang, and A. Halevy, "Clustering Query Refinements by User Intent," Proc. the 19th Int'l Conf. World Wide Web (WWW '10), Raleigh, NC, 2010.

[13] Y. Zhang, and G. Xu "Using Web Clustering for Web Communities Mining and Analysis, " 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, 2008

[14] J. Zhang, P. Zhao, L. Shang, and L. Wang, "Web Usage Mining Based On Fuzzy Clustering in Identifying Target Group", International Colloquium on Computing, Communication, Control, and Management, Vol. 4, Pp. 209-212, 2009

[15] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press., vol 1, pp. 281–297. 1967

[16] K. J. Kim, and H. Ahn, "A recommender system using GA K-means clustering in an online shopping market," Expert Systems with Applications, vol. 34, pp. 1200-1209, 2008

[17] M.P. Yadav, M. Feeroz, and V.K. Yadav, "Mining the customer behavior using web usage mining in e-commerce," Third International Conference on Computing Communication & Networking Technologies, Coimbatore, India, 2012

[18] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro, "The Intention Behind Web Queries," Lecture Notes in Computer Science, vol. 4209, pp. 98-109, 2006

[19] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Effective Personalization Based on Association Rule Discovery from Web Usage Data," Proceedings of the 3rd international workshop on Web information and data management, New York, NY, 2001

[20] M. Dimitrijević, and Z. Bošnjak, "Discovering interesting association rules in the web log usage data," Interdisciplinary Journal of Information, Knowledge, and Management, vol. 5, pp. 191-207, 2010.

[21] R. Suguna, and D. Sharmila, " Association Rule Mining for Web Recommendation," International Journal on Computer Science and Engineering, vol. 4, no. 10, pp. 1686-1690, 2012

[22] V. Nebot and R. Berlanga, "Finding association rules in semantic web data," Knowledge Based System, vol. 25 pp. 51-62, 2012

[23] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487-499, Santiago, Chile, 1994

[24] B.S.Kumar, and K.V. Rukman, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms," International Journal of Advanced Networking and Applications, vol. 01, i. 6, pp. 400-404, 2012.

[25] S. Veeramalai, N. Jaisankar, and A. Kannan, "Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy", International journal of computer science & information Technology, vol.2, no.4, pp. 60-74, 2010

[26] R. Mishra, and A. Choubey, "Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data," International Journal of Computer Science and Information Technologies, vol.3, (4), pp. 4662-4665, 2012

[27] http://en.wikipedia.org/wiki/Web_mining

[28] http://www.tutors-live.com

[29] http://tutor.104.com.tw/dsp_pa_list.cfm