Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 45 (2015) 60 – 69

Procedia Computer Science

International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

Dynamic Recommendation System Using Web Usage Mining for E-commerce Users

Prajyoti Lopes[1], Bidisha Roy[2]

[1,2]Department of Computer Engineering, St. Francis Institute of Technology, Mumbai, Maharashtra 400103, India
prajyotilopes@gmail.com
bidisha.bhaumik.roy@gmail.com

Abstract

E-commerce organizations are growing exponentially with time in terms of both business and data. Many organizations rely on these websites to attract new customers and retain the existing ones. In order to achieve this goal web log files can be used that records customer's access patterns. Using traditional web usage mining techniques in an enhanced manner valuable patterns and hidden knowledge can be discovered. This paper focuses on providing real time dynamic recommendation to all the visitors of the website irrespective of been registered or unregistered. Action based rational recommendation technique is proposed that makes use of lexical patterns to generate item recommendation. Effectiveness of the proposed system is evaluated by collecting real time E commerce data and comparing the system with user based and product based techniques. Results prove that the proposed system yield good quality accuracy and minimizes limitations of traditional recommendation system.

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

Keywords: Common log file; E-commerce; Lexical Patterns; Personalized recommendation; Web usage mining

1. Introduction

Today number of internet users using web to perform day to day transactions are increasing. A remarkable number of companies are encroaching towards internet to sell their products and services [1]. This revolution towards E-commerce has changed, conventional way of doing businesses. This rapid expansion has resulted in, new

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).
doi:10.1016/j.procs.2015.03.086

challenges to both companies as well as customers. Customers are overloaded with multiple choices for a specific product, which results in a confused and lost state. It has become trivial for the webmasters to evaluate whether the products and services provided are catering to the needs of the customers or not [2]. Therefore it is important to devise new marketing strategies such as one-to-one marketing and customer relationship management [3]. One effective solution to handle this issue is to provide personalized recommendation to individual user i.e. providing the customer with the type of product recommendation list he or she is interested in. A promising solution to overcome this issue is recommendation system which provides and guides the customer with type of product he or she is interested in buying/purchasing. Till date a wide variety of recommendation system have been devised and implemented. Recommendation systems can be broadly classified into two: Content-based and Collaborative filtering systems. Content based system takes into consideration the product attributes to generate recommendation. Collaborative filtering system makes use of customer - product interaction and ignores the other facts to provide recommendations [4] [5]. Despite significant progress and widespread use these recommendation system suffers from following limitations. The first one focuses on scalability. As number of consumers and products increases in number it slows down neighbor selection per second. Another issue that can be pinpointed is most of recommendation systems make use of binary transaction (clickstream) data. i.e. whether a specific item is purchased or not. However many a times they are unable to exploit intrinsic characteristics of these data that can be used to provide better recommendation [4]. Another drawback of conventional recommendation system is retargeting i.e. providing the customer with the same product that is already purchased. Studies in [6] indicate that web usage mining can act as an effective solution to overcome the limitation of traditional recommendation systems.

1.1. Web Usage Mining:

Web usage mining focuses on predicting users' preferences and behavior by analyzing web logs with help of traditional data mining techniques [2]. Customer's clickstream data can act as a very rich source of information. Clickstream indicates user's path through a website. Clickstream data is captured and maintained in web log files. Strategic use of navigational data can be very helpful in providing effective recommendation. Good quality recommendation systems will not only help in satisfying customers preferences for a product but also in improving sales and attracting new consumers. Indigent quality of recommendation, results in two types of peculiar errors, false negatives: these are the items not recommended even though the customer likes it. False positive these are the items recommended even though the customer dislikes it. In an E-commerce domain the most important errors that need to be handled and circumvented are false positives errors, which can result in unsatisfied customers and minimize their possibility to revisit the site once again.

This paper proposes a rational based recommendation technique that makes use of usage data and integrates it with user, sales data to provide better recommendations. Our research tries to provide effective recommendation to all visitors of an E-commerce site. The rest of the research work is organized as follows section II provides with an overview of data preprocessing and data mining techniques in use. Section III provides details of proposed system. Section IV focuses on results and discussion. In section V we conclude the paper along with future scope.

2. Related work

Cooley and et.al defined the term web usage mining for the first time and aims on predicting user's preferences and behavior [7]. The entire web usage mining process is divided into three phases namely data preparation, pattern discovery and pattern analysis as discussed by in [3]. Most of the data needed for web log analysis resides on web servers, proxy servers, enterprise logs, web clients etc. Many a times these data is ambiguous in nature and needs to be cleaned for further analysis. In order to yield better recommendation results good quality of data should be served as an input. Evaluation conducted in [8] [9] show that 80% of time spend in data mining is consumed in log data preprocessing. Web log data has peculiar characteristics and therefore detailed description about various steps followed is given in [9] [10]. The pattern discovery tasks include the discovery of association rules, sequential patterns, user classifications etc. In [11] explains Naïve Bayesian techniques for dynamic mining of user's interest

navigation pattern. However this technique is time consuming. In [12] explain agglomerative hierarchical clustering method that makes use of Euclidean distance to calculate similarity measure and clusters the users having similar browsing feature. In [3] makes use of association rule mining, product taxonomy and web usage mining to provide personalized recommendation. It tries to overcome certain limitations of collaborative filtering. We propose a system that will try to overcome limitations of traditional recommendation systems.

3. Proposed work

In the proposed system users interacts with web portal and users click stream data is maintain in raw log file. Multiple preprocessing and data cleaning task are performed to extract valuable information from raw log files and transform it in structured form. This cleaned data is further used to discover patterns, hidden rules and provide top n product recommendation to all users of E-commerce site. Fig. 1. Illustrates detailed description of the proposed work.

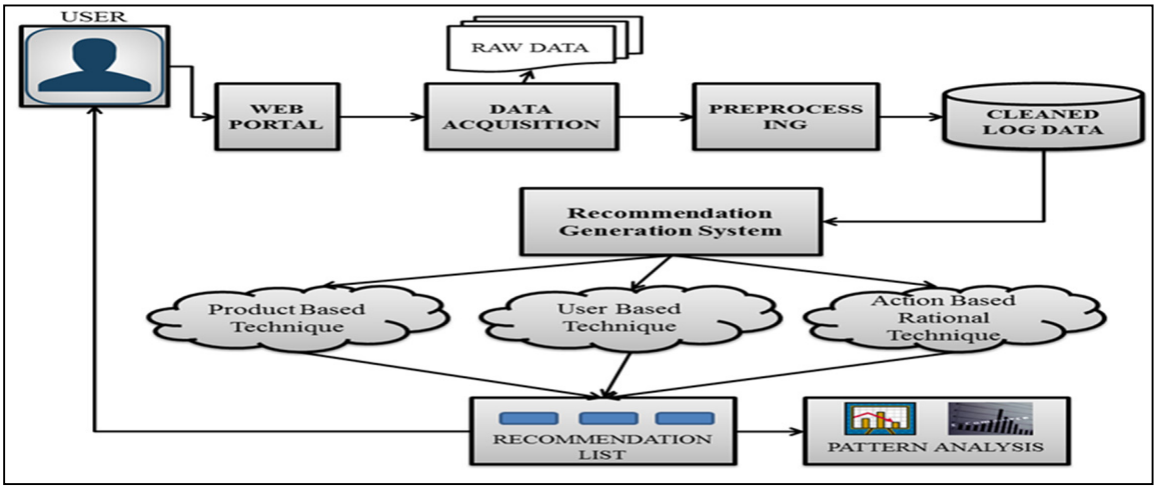

Fig. 1. Proposed Recommendation System

3.1. Data Acquisition

In this phase the entire navigational data which includes all the web pages visited is collected and stored. The proposed work makes use of common log file format to maintain the data, important attributes namely IP address, timestamp, status code, URL, method (GET and POST), user agent and Referrer URL are recorded and used for further analysis. This data obtained is highly unstructured and inconsistent in nature. And therefore has to be preprocessed for further analysis.

3.2. Data Preprocessing:

Good quality input data needs to be served for better analysis. In this phase inconsistent, redundant data is eliminated using following steps as depicted in Fig. 2. Field separation stage focuses on distinguishing one attribute from another by making use of separator character such as space. In data cleaning stage we filter out outliers data. We check for URL suffixes. Log entries having filename suffixes such as gif, jpeg, tif, jpg are discarded. All records having failed http status code are removed i.e. status code greater than 200 and less than 299 are eliminated. In user differentiation phase we assign unique user ID, to each IP address and registered users to differentiate one customer from other. Finally we construct session in session identification phase. In session clustering phase we group

together session belonging to unique user. Session provide us with complete set of activities done by the user in specific time period. Finally in data formatting stage we place the data in tabular form.

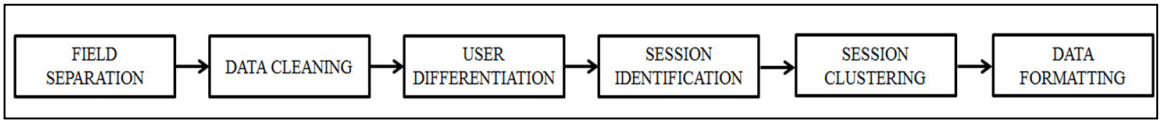

Fig. 2. Data Preprocessing Steps

3.3. Recommendation system

Three different recommendation systems have been proposed. Product based recommendation technique works specifically for unregistered users. Dependent on users IP address session are constructed. Based on session details recommendation list is generated. Combination of session is taken to generate final recommendation list. User based technique is used for registered users. In this technique based on user's navigational details appropriate recommendations are provided. A detailed description is given in [13]. For all the three approaches customer inclination analysis is done dependent on three parameters: clicked through, basket placement and purchased product.

3.3.1 Action based Rational Recommendation Technique

This technique provides dynamic recommendation as per changing user behaviour. It constructs lexical patterns by converting sequence of products visited into meaningful character string. This technique is suitable for all users (registered as well as unregistered). The entire technique is divided into two steps:

Step 1: Capturing traversal patterns and converting to lexical patterns:

In this approach for each session s we construct an object. This object maintains information for all products in the session such as product name (p_id), category name (c_id), frequency (f) which indicates number of times product visited and time spent on page (tsp). Fig. 3. Illustrates an example of lexical string.

"31.33":{"page_name":"31.33","product_id":31,"time_spend":0.52,"count":2},"49.57":{"page_name":"49.57","product_id":49,"time_spend":0.73,"count":1},"47.18":{"page_name":"47.18","product_id":47,"time_spend":0.4}

Fig. 3. Traversal String Pattern

Details highlighted in red indicate 31.33 is your object name where 31 is p_id and 33 is c_id , 0.52 is page stay time (tsp) and 1 indicates frequency of visit to that product in a particular session.

Step 2: Recommendation generation method:

In this technique based on user's category (whether registered or unregistered) corresponding product recommendation is made. For unregistered user IP address is the deciding factor and for registered user unique user id is used. In this technique we fetch session based on most recent timestamp (ts). We consider below parameters before placing product in recommendation list as given in Equations 1 to 3.

User Interest Measure (μ): It implies whether a user is actually interested in a specific product. μ is calculated using page stay time.

$$\textbf{User Interest Measure } (\mu) = \begin{cases} 1, \text{ if } tsp \geq 0.05 \\ 0, \text{ if } tsp < 0.05 \end{cases} \quad (1)$$

Frequency (f): Highest frequency product is placed in list. It two products have same frequency value then we check for tsp. Therefore if $f.p_1 > f.p_2$ place p_1 product in recommendation list.

$$\textbf{Wish list buffer } (\eta) = \begin{cases} 1, \text{ if present in wish list, add in recommendation list} \\ 0, \text{ if not present} \end{cases} \quad (2)$$

Based on above parameters we construct initial recommendation list and then fetch related products based on category and manufacturing details to produce final list of products. We also calculate similarity measure to determine which lexical string is to be retrieved in case if multiple session have same product.

$$\textbf{Similarity Measure} = \frac{\text{max of products in session (s1, s2)}}{\text{total number of products in each session}} \quad (3)$$

The working of entire technique is given in below Fig. 4. In this approach dependent on the type of customer (whether registered or unregistered) corresponding technique is involved. Depending upon the current traversal pattern recommendation list is generated. Different parameters namely page stay time, count, wish list buffer content and similarity measure corresponding product recommendation string is generated. Dynamic recommendation list is generated as per changing behavior of the user in the current session.

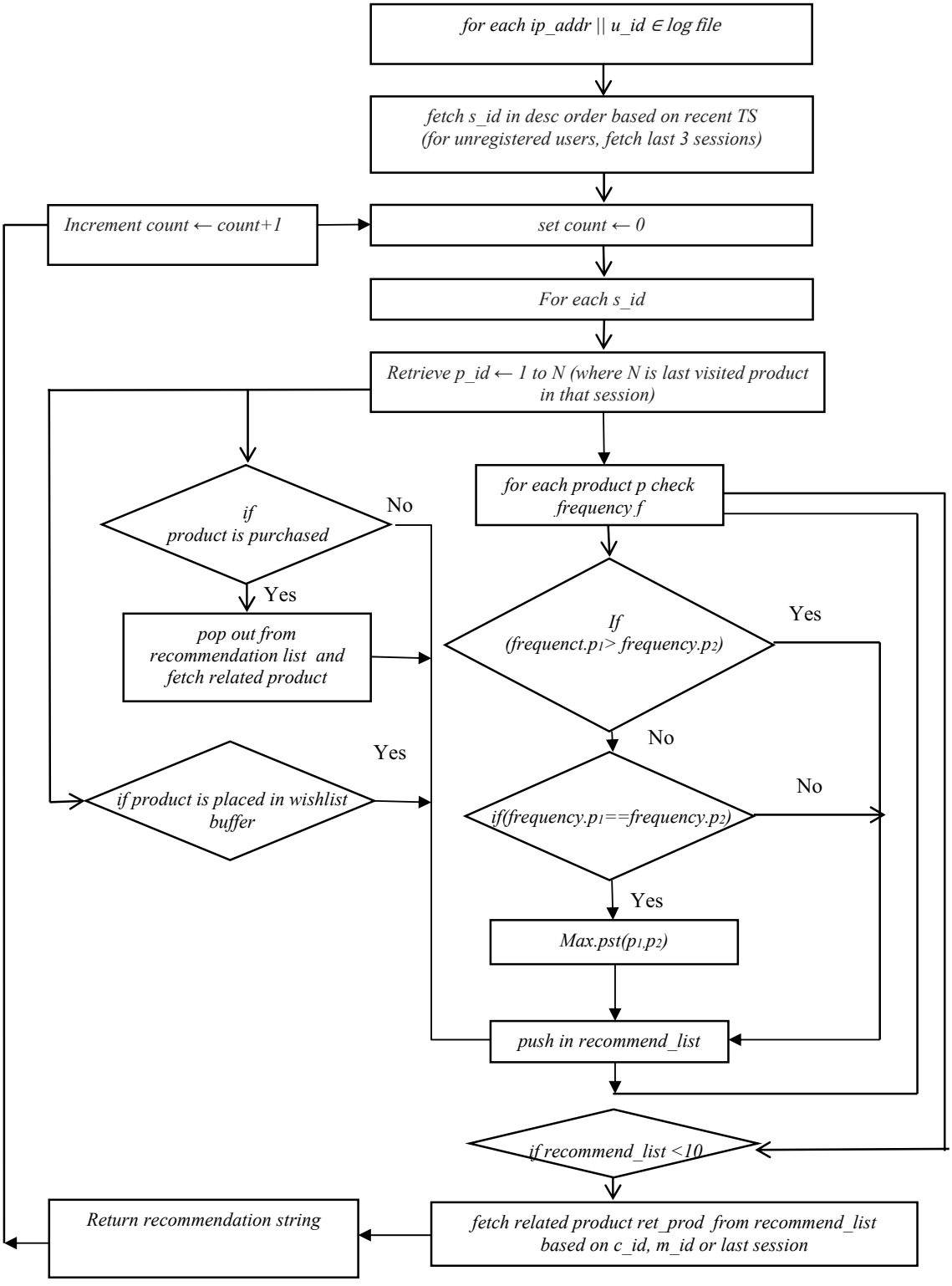

Fig. 4. Action Based Rational Recommendation Technique

4. Results and discussion

The proposed system is implemented using XAMPP server, phpMyAdmin and Sublime Text 3 IDE. Opencart [14] which is an MVC based open source shopping cart system was used to write the testing program. For validating our proposed system, web portal was developed that offered different electronic products for the customers. The database used was real time data containing 1121 records from web portal "pjshopping.asia". Fig. 5. Illustrates the GUI of the web portal offering multiple categories of electronic product. In Fig. 6. A sample of uncleaned log file is shown that contains navigational data. In Fig. 7. Cleaned log file in appropriate structured form is shown.

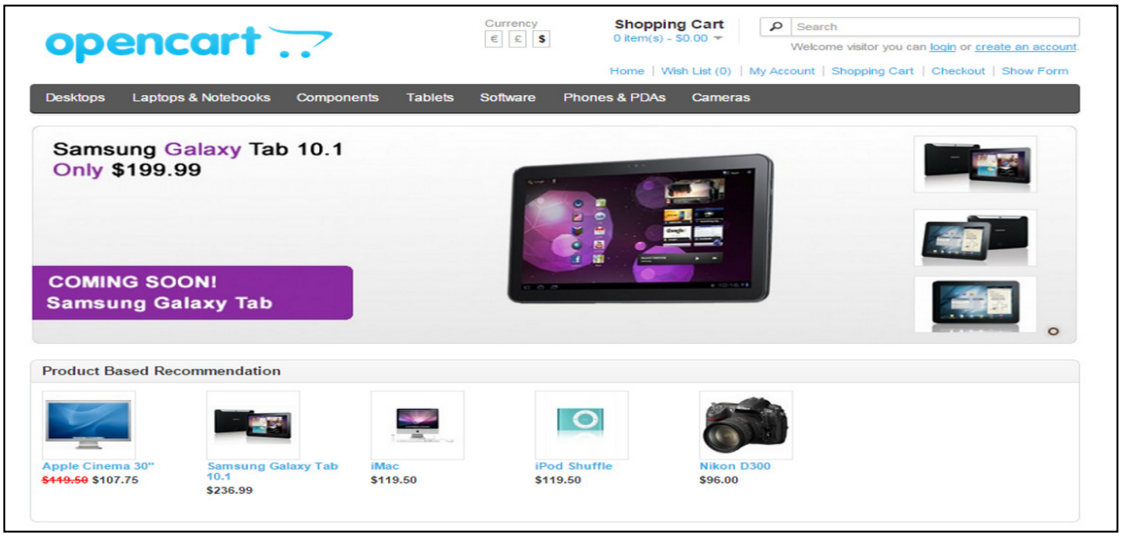

Fig. 5. GUI of Proposed System

```
1  114.143.229.227---2014-11-28 14:47:53---GET---http://pjshopping.asia/index.
   php?route=product/product&amp;product_id=48---HTTP/1.0---1024603552---http://
   pjshopping.asia/---Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like
   Gecko) Chrome/39.0.2171.65 Safari/537.36---0.08---48
2  114.143.229.227---2014-11-28 14:48:58---GET---http://pjshopping.asia/index.
   php?route=common/home---HTTP/1.0---501572755---http://pjshopping.asia/index.
   php?route=product/product&amp;product_id=48---Mozilla/5.0 (Windows NT 6.1)
   AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.65 Safari/537.36---1.08---0
3  114.143.229.227---2014-11-28 14:51:10---GET---http://pjshopping.asia/index.
   php?route=product/product&amp;product_id=42---HTTP/1.0---432125798---http://
   pjshopping.asia/index.php?route=common/home---Mozilla/5.0 (Windows NT 6.1)
   AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.65 Safari/537.36---1.05---42
4  114.143.229.227---2014-11-28 14:55:55---GET---http://pjshopping.asia/---HTTP/1.0---
   1054415468---http://pjshopping.asia/---Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.
   36 (KHTML, like Gecko) Chrome/39.0.2171.65 Safari/537.36---0.42---0
5  114.143.229.227---2014-11-28 14:56:16---GET---http://pjshopping.asia/---HTTP/1.0---
   1157146850---http://pjshopping.asia/---Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.
   36 (KHTML, like Gecko) Chrome/39.0.2171.65 Safari/537.36---0.35---0
6  114.143.229.227---2014-11-28 14:56:28---GET---http://pjshopping.asia/index.
   php?route=product/product&amp;product_id=42---HTTP/1.0---1162873841---http://
   pjshopping.asia/---Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like
   Gecko) Chrome/39.0.2171.65 Safari/537.36---0.2---42
```

Fig. 6. Uncleaned Log File

id	request	date	file_path	referrer	session_id	ip	server_protocol
1	GET	2014-11-28	data/demo/ipod_classic_1.jpg	http://pjshopping.asia/	1	114.143.229.227	HTTP/1.0
2	GET	2014-11-28	http://pjshopping.asia/index.php?route=product/pro...	http://pjshopping.asia/	1	114.143.229.227	HTTP/1.0
3	GET	2014-11-28	http://pjshopping.asia/index.php?route=common/home	http://pjshopping.asia/index.php?route=product/pro...	1	114.143.229.227	HTTP/1.0
4	GET	2014-11-28	data/demo/apple_cinema_30.jpg	http://pjshopping.asia/index.php?route=common/home	2	114.143.229.227	HTTP/1.0
5	GET	2014-11-28	http://pjshopping.asia/index.php?route=product/pro...	http://pjshopping.asia/index.php?route=common/home	2	114.143.229.227	HTTP/1.0
6	GET	2014-11-28	http://pjshopping.asia/	http://pjshopping.asia/	3	114.143.229.227	HTTP/1.0
7	GET	2014-11-28	http://pjshopping.asia/	http://pjshopping.asia/	3	114.143.229.227	HTTP/1.0

Fig. 7. Structured Log Data

4.1. Quality Evaluation:

Accuracy is the parameter used to evaluate the effectiveness of proposed system with respect to all three techniques. A matrix is constructed to measure accuracy as shown in Table 1 below.

Table 1. Recommendation Matrix

	Recommended Items by the System	Items Not Recommended by the System
Expected Item	True Positive(TP)	False Negative(FN)
Not an Expected Item	False Positive(FP)	True Negative(TN)

Based on the recommendation matrix we calculate recall, precision and accuracy as shown in Equations 4 to 6.

Recall can be defined as a fraction of all relevant items that are recommended by the system.

$$Recall = \frac{True\_Positive(TP)}{True\_Positive(TP) + False\_Negative(FN)} \quad (4)$$

Precision is a fraction of all the recommended products that are relevant.

$$Precision = \frac{True\_Positive(TP)}{True\_Positive(TP) + False\_Positive(FP)} \quad (5)$$

The Accuracy is ratio of true positives to sum of all the items recommended

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Accuracy of the system was measured by collecting feedback from the customers based on the recommendation matrix. Fig. 8. Depicts the values of Average Precision, Average Recall and Average Accuracy obtained after the customer's feedback.

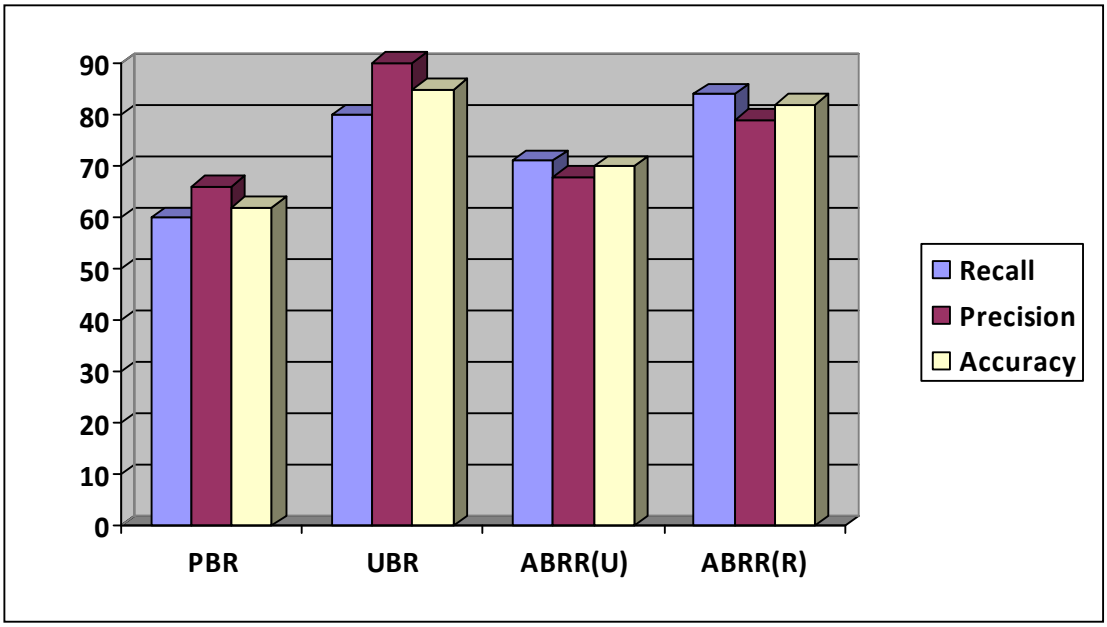

Fig. 8. Accuracy Measure

Where PBR is product based recommendation, UBR is user based recommendation; ABRR (U) is action based rational recommendation for unregistered users and ABRR (R) is action based rational recommendation for registered users. Results show that ABRR (U) gives us an accuracy of approximately 70 percent and ABRR (R) provides an accuracy of 82 percent.

5. Conclusion

In this research work we focus on providing good quality product recommendations to all the users especially unregistered ones of E-commerce site. The beauty of the proposed system is it dynamically provides recommendation as per changing users' behavior and traversal patterns by making use of web usage mining and constructing patterns from the historical data. The proposed recommendation system minimizes the false positive errors that occur frequently in traditional recommendation system. Also issue of binary ratings and cache memory are handled by the system thereby providing good quality recommendations. Results prove that accuracy of approximately 80 to 85 percent is achieved for registered user and 65 to 70 percent for unregistered user in rational recommendation technique, which is better than product based technique and almost equivalent to user based approach. The recommendation system has the potential to attract new customers and retain existing ones. This technique can help the E-commerce organization have competitive edge in the market and can be helpful in forecasting demands and sales for a specific product. It would be interesting to evaluate the proposed technique with different conventional recommendation approaches and measure its accuracy. This proposed system can also be tested for other application areas like movie recommendation, music recommendation etc.

References

[1] Y. Cho, and J. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce", *Expert systems with Applications*, vol. 26, no. 2, pp. 233-246, February 2004.

[2] Q. Song, and M. Shepperd, "Mining web browsing patterns for E-commerce", *Computers in Industry*, vol. 57, no. 7, pp. 622-630, 2006.

[3] Y. Cho, J. Kim, and S. Kim, "A personalized recommender system based on web usage mining and decision tree induction", Expert Systems with Applications, vol. 23, no. 3, pp. 329-342, October 2002.

[4] Z. Huang, D. Zeng, and H. Chen, "A comparative study of recommendation algorithms in e-commerce applications", *IEEE Intelligent Systems* vol. 22, no. 5 pp. 68-78, 2007.

[5] J. Lee, M. Sun, and G. Lebanon, "PREA: Personalized recommendation algorithms toolkit." *The Journal of Machine Learning Research, vol.*13, no. 1, pp. 2699-2703, 2012.

[6] B. Mobasher, R. Cooley, J. Srivastava, "Automatic personalization based on Web usage mining" , *Communications of the ACM*, vol.43 no. 8 pp. 142-151, 2000.

[7] Y. M. Huang, Y.H. Kuo, J.N. Chen, and Y.L. Jeng, "NP-miner: A real-time recommendation algorithm by using web usage mining", *Knowledge-Based Systems*, vol.19, no.4, pp. 272-286, 2006.

[8] C.R. Varnagar, N.N. Madhak, T. M. Kodinariya, and J. N. Rathod, "Web usage mining: A review on process, methods and techniques", *Information Communication and Embedded Systems (ICICES), International Conference on. IEEE*, pp. 40-46, 2013.

[9] P. Nithya, and P. Sumathi, "Novel pre-processing technique for web log mining by removing global noise and web robots*." In Computing and Communication Systems (NCCCS) IEEE*, pp. 1-5,2012.R. Cooley, B. Mobasher, J. Srivastava, "Data preparation for mining world wide web browsing patterns", *Knowledge and information systems,* vol.1, pp. 5-32, 1999.

[10] M. Khosravi, and M. J. Tarokh, "Dynamic mining of user's interest navigation patterns using naive Bayesian method." *In Intelligent Computer Communication and Processing (ICCP), IEEE International Conference on*, pp. 119-122, 2010.

[11] B. Devi, Y. Devi, B. Rani, and R. Rao, "Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce." *Procedia Engineering*, vol. 30, pp. 20-27, 2012.

[12] P.Lopes, and B. Roy, "Recommendation System using Web Usage Mining for users of E-commerce site", *International Journal of Engineering Research & Technology,* vol. 3, issue 7, 2014.

[13] "OPENCART", [ONLINE].Available:http://www.opencart.com, 2013.