Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 12, December -2017

ONTOLOGY BASED PERSONALIZATION FOR WEB SEARCH IN DATA MINING

S.Kannan[1], Dr.G.N.K. Suresh Babu[2]

[1]*Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore, India.*
[2]*Professor, Department of Computer Applications, Acharya Institute of Technology, Bangalore, India.*

ABSTRACT:- *Web usage mining is the kind of data mining application to extract the knowledge from the web.Nowadays the internet growing lots of web pages which is developed. The most important researcher is focused on web based model. Web Personalization is one of the areas of the Web usage mining that can be used to deliver the dynamic content to the particular user or segments of users. The goal of web personalization using web usage mining is to identify interesting patterns from web usage data and recommend objects to the user which consists of products, text, links and so forth. In this paper, web personalization process which is focused on user profiling based on ontology. This model providesinterest score which is based on their user's interestand it can achieve more accurate personalization.*

Keywords*: web personalization, user profiles, ontology, data mining, interesting score.*

INTRODUCTION

Data mining is evolved as one of the most dynamic, vibrant and active area of computer science engineering. The wide field of classical data mining contains the extension for web based environment in form of web mining. Day by day, the web is expanding by millions of web pages, adding to the already developed millions of web pages already. Because of its rapid and massive growth, the resulting pages of information are mismanaged in terms organization and structure. In this information overloaded scenario that continues to expand users often feels disoriented and lost its focus. Thus we need better techniques for web based data. The web mining is an approach of searching, discovering and retrieving the user oriented and interesting information and patterns from a huge dataset [1]. This is one of the emerging fields in the industry.

Web Usage Mining 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. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Some of the typical usage data collected at a Web site includes IP addresses, page references, and access time of the users. It is also called web log mining. The web usage mining is a process of retrieving pattern from web access log. The three processes which are used in Web usage mining is Preprocessing, Pattern discovery & analysis [2].

Personalization requires the different goals and also it is useful to develop different business application. That delivery of content tailored to a particular user.Personalization can be extended to recommend news/blogs related to a user's interests; augment the recent document list in document readers using favorite document lists; identify users with common interests for collaborative recommendation; use geographical location and information from the web to suggest events in the city that match a user's interests etc.Personalizing web pages is one of the efficient parts of web-mining, which enables to understand user interests to offer services and enables them to discover web pages, text documents, multimedia files, images and other types of resources from web according to their choice[3].

Personalization is a methodology of learning a user profile of user modeling from his browsing styles. With using the technique of personalization, the Web user would prefer an intelligent Web server which capable to learn their information needs and preferences. The Web provides a direct interaction medium between the vendors of products and services, and their customer with very effective or low cost. Web personalization has numerous applications. The goal of personalization systems is to provide users with what they need or want asking the users preferences [4].

Web personalization is a strategy, a marketing tool, and an art. The objective of a Web personalization system is to "provide users with the information they want or need, without expecting from them to ask for it explicitly" [5]. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. A personalization mechanism is based on explicit preference declarations by the user and on an iterative process of monitoring the user navigation, collecting its requests of ontological objects and storing them in its profile in order to deliver personalized content [6].

LITERATURE SURVEY

As web mining is a very broad topic presenting various areas for research purpose, technology and development but the area of our concern is web personalization based on web usage mining where Web log data collected from server can be used to select and extract interesting patterns. As this area is fresh and a lot of development is required but still appreciable and tremendous work is done by people.

"Jeh and Widom" [7] proposed a method for personalized web searching by applying modifications in the global Page-Rank algorithm. Instead of starting from random pages on the web, they have used the concept of bookmarks which user saves generally during navigation over internet.

Probability LatentSemantic Analysis (PLSA) is an efficient approach to capture the latent or hidden semantic relationships and knowledge among the co-occurrence activities, the system characterize the web user segments and provide dynamic and personalized recommendations. The Markov model and click-stream tree concept which are combined in and a hybrid recommendation model is designed for web users and it recommends web pages. Differentially private RS was proposed in and the model guarantees the privacy of the web users.

Spiliopoulou et al. (2002) [8] study performance of various such heuristic techniques. Server and proxy logs are also inaccurate in capturing temporal aspects of user interactions. Timestamps recorded in these logs for page requests incorporate network transfer time. Due to nondeterministic behavior of the network, there is no trivial way to filter out these noise data. On the other hand, if temporal features are captured on the client side, occurrence times of all user interactions can be recorded as exactly as required. Thus, data collected at the client side provide us with the most accurate spatial and temporal information about user interactions with the web.

Mobasher et al. [9] utilize Web transaction and page clustering techniques, which is employing the traditional k-means clustering algorithm to characterize user access patterns for Web personalization based on mining Web usage data. These proposed clustering-based techniques have been proven to be efficient from their experimental results since they are really capable of identifying the intrinsic common attributes revealed from their historic clickstream data. Generally, these usage patterns are explicitly captured at the level of user session or page. They, however, do not reveal the underlying characteristics of user navigational activities as well as Web pages. For example, such discovered usage patterns provide little information of why such Web transactions or Web pages are grouped together, and latent relationships among the co-occurrence observation data have not been incorporated into the mining processes as well. Thus, it is necessary to develop LSA-based approaches that can reveal not only common trends explicitly, but also take the latent information into account implicitly during mining.

Zhang, Lakshmanan, and Zamar[10] studies the problem of extracting data records from a large set of Web pages. What is interesting about this paper is that it proposes a novel method to estimate the current coverage of the results of the system, based on capture-recapture models with unequal capture probabilities. Techniques are also proposed to estimate the error rate of the extracted information. To evaluate the method and ideas proposed in this paper, a large number of experiments have been conducted to demonstrate their effectiveness.

Chang, He, and Zhang [11] studies dynamic "onthe-fly" semantics discovery for large scale integration on the "deep Web," and proposes "holistic mining" as a conceptual framework unifying initial works and a general approach for such large-scale integration. It reports three sample tasks as evidences: interface extraction, schema matching, and query translation. To generalize, it then proposes holistic mining as a unified insight to observe and leverage "hidden regularities" across holistic sources for large scale integration, and outlines future challenges.

To analyze the web data, various approaches have been proposed by researchers across the globe in past few years: Magdalini Eirinaki et al. [12] classified these approaches into four major classes: Content-based filtering systems, Social or Collaborative filtering systems, Manual rule-based filtering systems, and Web usage mining based systems. Content-based filtering systems model the behavior of individual user based on his past interests, personal preferences, and browsing behavior. Once user modeling is completed, system starts recommending items (to users) that match individual user's profile.

Ronaldo Lima Rocha Campos et al., 2011 [13] proposed a Multiagent based system application model for indexing, retrieving and recommendation learning objects that are stored in different and heterogeneous repositories. In order to improve the accuracy, they have come up with an information retrieval model, which is based upon the multi-agent system approach and an ontological model.A challenging problem in recommendation systems deals with unvisited or newly added pages. Rana Forsati et al., 2009 [14] addresses this problem by introducing a novel Weighted Association Rule mining algorithm. This method can improve the overall quality of web recommendations.

Sieg et al., [15] review an approach to personalized web search which incorporates models of "user context" based on ontological profiles by assigning implicitly derived user's interest scores to concepts in the prebuilt domain ontology. They define "context" as the immediate and past user activities as well as knowledge from a pre-existing ontology as an explicit representation of the domain of interest.

E-learning environments have gained a wide acceptance and are currently being used by a wide variety of students with different skills, backgrounds, interaction preferences, and learning styles. Such diversity has brought new opportunities

and new challenges [16]. As e-learning systems evolve toward more complex configurations, modeling context becomes more important either to personalize or to reuse that contextual information [17]. Therefore, it is important to discuss ways to improve these environments by considering the students' context. This is the reason why we decided to test our approach in this domain. Our intention is to increase even more the personalization capacities of actual e-learning systems by making use of ontologies to model the context of students in different scenarios. Also, we aim to test our proposal's ability to learn the relevant context for the student in concrete e-learning situations.

PROBLEM IDENTIFICATION

1. The Problem is scalability, data scarcity, Cold-star problem which affect coverage and poor quality recommendationsthose users who have rated small number of items. The cold-start problem in recommender systems where no initial information is available early on upon which to base recommendations

2. Server and proxy logs are also inaccurate in capturing temporal aspects of user interactions while timestamps recorded in these logs for page requests incorporate network transfer time.

3. A challenging problem in recommendation systems deals with unvisited or newly added pages. Need to improve the overall quality of web recommendation. User profile can be built over a period of time, until the system gets a minimum level of data, it may not provide the recommendations till such level

4. Too much filtering can cause a performance issue which in turn leads the user to lose the interest.

RESEARCH METHODOLOGY

The proposed research methodology is based on recommendation model for web personalization in web usage mining. The ontology based technique is used for recommendation to the users. The below figure architecture shows the proposed flow.

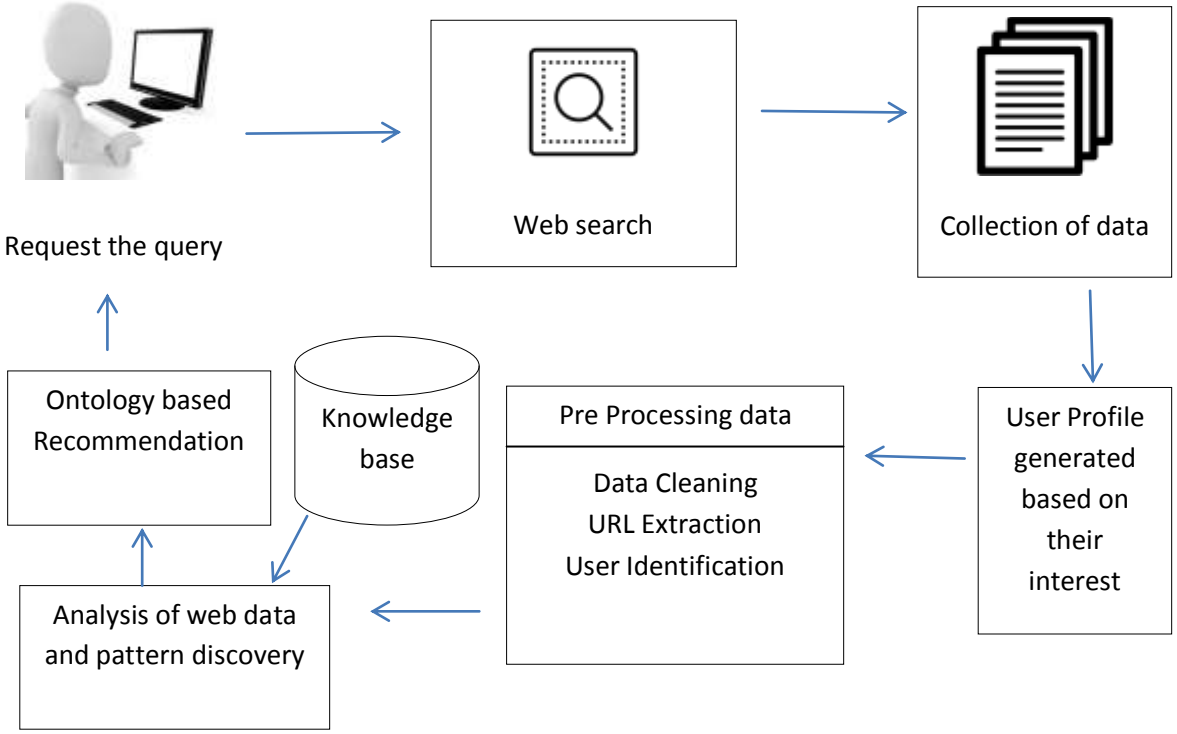

Figure proposed architecture flow.

Web personalization serves the purpose up to some extent by understanding the user however, relevant results cannot be provided to the user With-out understanding the context of the user. Web personalization has four phases, namely:

a) Collection of web data
b) Web data Pre processing
c) Web data analysis and pattern discovery
d) Recommendations phase

Collection of web data: Implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules.

Web data preprocessing: Preprocessing may include cleaning data, user identification, URL extraction, user pattern extraction.

a) ***Data cleaning:*** The data cleaning process means removing the inconsequential data from database log. This process is used to reduce the size of data, if a request sent for a web page there are various attributes called and many contents loaded for that particular request. The result can be image and graphics, as we loaded for the HTML tags as our requirement, we only need data which is generated by user, and it is necessary to make sure that user request exist in server logs. So task which is generated by user should be ignored and remove from log files.

b) ***User Identification:*** This step is after the data cleaning step. This is the main task for data preprocessing process. In this must identify the unique users from data log. If data log shows many user having same IP address, by tracking the agent in which the request have been produced. Still differentiate for obtaining the different user sessions. If two users have common IP address then both of users considered as a single user.

c) ***URL Extraction:***After identified the user**.** The URL can beextracted from the web server.Identified the unique URL and assign the unique ID to each URL.

Analysis of web data and Pattern discovery: Also known as Web Usage Mining, this step applies machine learning or Data Mining techniques to discover interesting usage patterns and statistical correlations between web pages and user groups.Pattern discovery aims to detect interesting patterns from the preprocessed web usage data i.e. mining the data. It includes methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Pattern analysis step is also referred as recommendation phase in case of web personalization using web usage mining, where various URLs are recommended.Extracting useful patterns and rules using data mining techniques in order to understand the usersbehavior.

Page rank: - Page rank work by calculate the number of quality of link to a page for presumes how important a link or website. The definition of page rank is give: - we suppose a page has n pages (t_1…….t_n) and the parameter d is damping factor (set from 0 to 1).

The rank of page is given by:

$$\Pr(A) = (1-d) + d\left(\sum_{i=1}^{n} \frac{\Pr(t_i)}{\Pr(t_i)}\right)$$

Where page rank Pr (A) can be calculated by a simple iterative algorithm and it is correspond to the principle eigenvector of normalized link matrix of web. N is the number of document we have, link matrix is M, where the Mij entry is 1/ni, if there is a link form document j to document i, then we can compute the page rank on the graph which is dominant eigenvector of matrix A.

Recommendation phase: The recommendation phased is based on ontology which carried out the existence knowledge which describes some domain typically called common sense knowledge domain. In web personalization process, the web resources (e.g. web pages) are generally recommended as an added function. The ontological user-profiles approach successfully addresses the cold-start problem. Initially there will be existing domain ontology. When a user fires a query, then the initial user behavior is matched with existing concepts in domain ontology and relationships between these concepts.

In proposed system the ontology based recommendation is used. These ontology based recommendation is Spreading activation algorithm. The basic concept behind Spreading Activation is that all relevant information is mapped on a user as query with a certain "activation level". Spreading activation is similar to artificial neural networks, except instead of individual neurons, the level of analysis is larger; for example concepts, documents, or web pages. When applying spreading activation to web ontologies the unit is an entity and the connections relations between them.

The basic process is that each node receives some initial activation representing a stimulus (e.g. web page visited, or name in an email message). Each activated node then passes on some activation to connected nodes, usually with some weighting of connections determining how much gets spread to each. This is then iterated, either for some fixed time, or until the activations become stable. Mathematically this looks like:

$$a_i = \gamma I_i + \mu f(\Sigma w_{ij} * a_j)$$

Here, a_i is the activation of node i, a_i the activation at the next iteration, I_i is the initial activation, w_{ij} is the weight in the connection between node i and node j, and λ and μ are constants.

The function f is there because some form of threshold, or nonlinear transformation (e.g. logistic) is often applied to the incoming activation for a node. A common value of wij is g/nj where nj is the number of outgoing connections from node j (fan out), and g is a 'gain' parameter.

PERFORMANCE ANALYSIS

The performance analysis of proposed system is calculated the following factors such as efficiency, accuracy, scalability, quality of recommendation.

Quality of Recommendation:To provide the proper user with appropriate recommendations, the personalization system must be able to provide high quality recommendations based on the data gathered on each user.

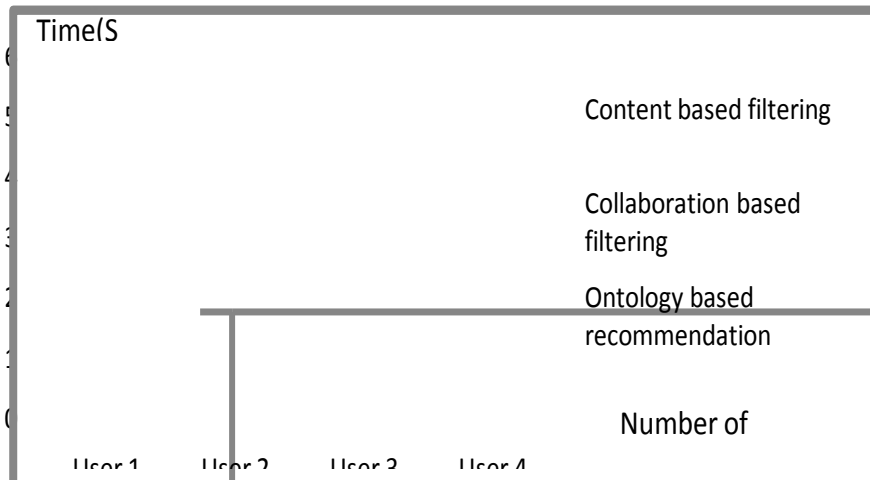

Figure 1 Quality of recommendationgraph

Accuracy: The personalized system can be recommendation based user interaction among the website which processed the large amounts of data. The accuracy of the recommended page become high while score the interesting pattern based on user profile.

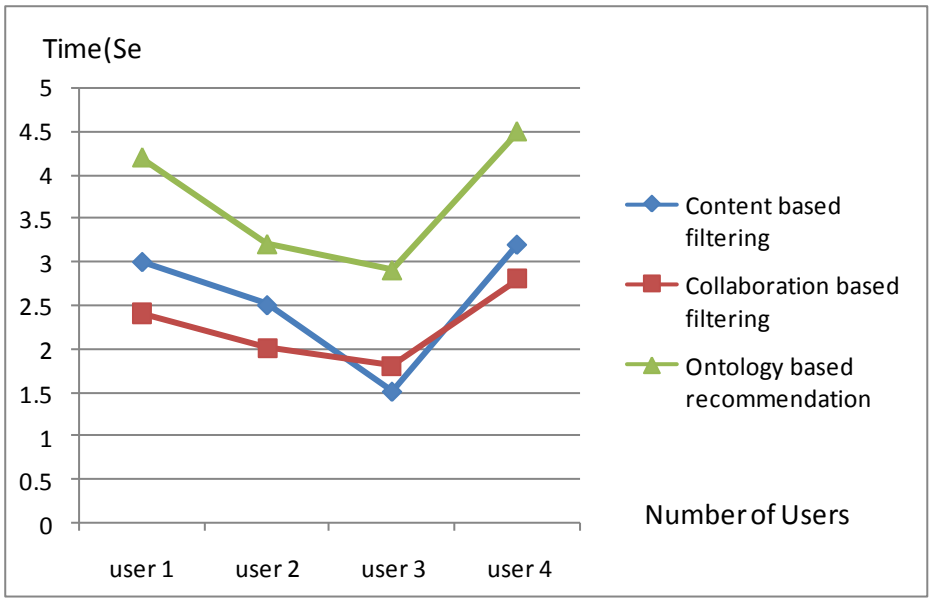

Figure 2 Accuracy level graph

Efficiency:An efficient web personalization system uses minimal processing in order to serve up the appropriate recommendation.Efficient processing will enhance end-user satisfaction by preventing slow responses in these real-time systems.

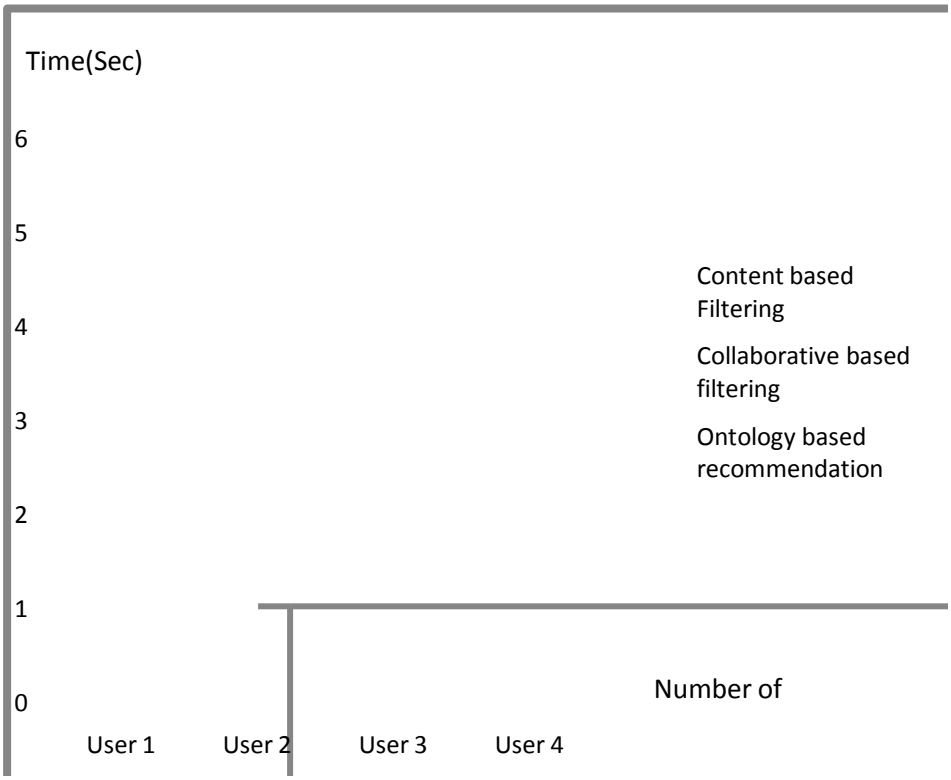

Figure 3 Efficiency graph

Scalability:A measure related to efficiency, system scalability ensures that the response time does not increase as additional data is collected by the personalization system.

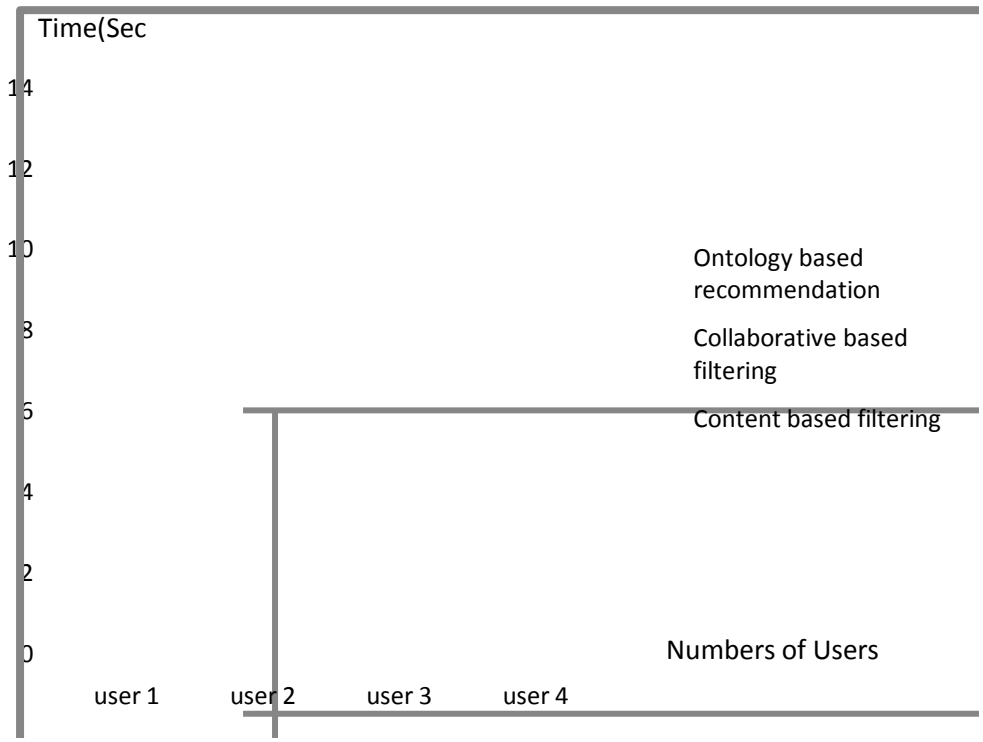

Figure 4 Scalability graph

Factors	Existing methods		Proposed method
	Content based filtering	Collaborative based filtering	Ontology based recommendation
Accuracy	3.2	3.5	4.9
Efficiency	3	2.4	3.5
Scalability	4.5	3.4	5
Quality of Recommendation	3.1	2.0	5.2

Table 1 for Existing system with proposed system

CONCLUSION

Web usagemining is useful for the pattern matching, sitereorganization, product/site recommendation etc.This paper can be focus on recommendation model that can be used for providing the particular product or item is recommended for user based on their interesting pattern. It can allow for more personalized interaction among the users. In future, web robot recommendation model can be used, that can be automatically recorded the user browsing information. The user whatever browsing in the internet that can be very much useful for web personalization in real time applications like e-commerce, amazon etc.

REFERENCE

1. J. Srivastava, O. Zaiane ed. Proceedings of,"WebKDD2002: Web Mining for Usage Patterns and User", 2002.
2. PurviDubey, Pramod S. Nair, Recommendation System for Web Mining: A Review, *International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 11, January 2015*
3. AnupamaPrasanth,Web Personalization Using Web Usage Mining Techniques, ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, Volume-3, Issue-3, 2016
4. VikasBhatnagar, Ruchi Davey, Contrasting Clustering Techniques for Web Personalization in Web Mining, IJCST Vol. 3, Issue 3, July - Sept 2012
5. M. Eirinaki and M. Vazirgiannis Athens University of Economics and Business, "Web Mining for Web personalization," *ACM Transactions on Internet Technology,* 2005.
6. D. Antoniou, M. Paschou, E. Sourla, and A. Tsakalidis, "A Semantic Web Personalizing Technique The case of bursts in web visits," presented at IEEE Fourth International Conference on Semantic Computing, 2010
7. G.Jeh, J. Widom,"Scaling personalized web search", In Proc. Intl. Conf. WWW, pp. 271–279, 2003
8. Spiliopoulou, M., B. Mobasher, B. Berendt, M. Nakagawa. 2002. Evaluating the quality of data preparation heuristics in web usage analysis. INFORMS Journal on Computing, to appear.
9. Mobasher, B., et al., Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Data Mining and Knowledge Discovery, 2002. 6(1): p. 61-82.
10. Zhang, Z., He, B., Chang, K. C.-C. Understanding Web Query Interfaces: Best-Effort Parsing with Hidden Syntax. SIGMOD-04, 2004
11. Chang, K. C.-C., He, B., Li, C., Patel, M., Zhang, Z. Structured Databases on the Web: Observations and Implications. SIGMOD Record, 33(3), 2004
12. Magdalini Eirinaki and Michalis Vazirgiannis, "Web Mining for Web Personalization," ACM Transactions on Internet Technology, Vol.3, No.1, 2003, pp.2-38
13. Ronaldo Lima Rocha Campos, Rafaela LunardiComarella and Ricardo AzambujaSilveira, "Multiagent Based Recommendation System Model for Indexing and Retrieving Learning Objects," Springer, Communications in Computer and Information Science Vol.365, 2013, pp.328- 339.
14. Rana Forsati, Mohammad Reza Meybodi and AfsanehRahbar, "An Efficient Algorithm for Web Recommendation Systems," In Proc. of IEEE/ACS International Conference on Computer Systems and Applications, May 2009, pp.579-586.
15. A. Sieg, B. Mobasher, and R. Burke, "Web search personalization with ontological user profiles," Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07, 2007, p. 525
16. Iglesias A, Martinez P, Aler R, Fernandez F (2009) Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning. Appl Intell 31:89–106
17. Eyharabide V, Gasparini I, Schiaffino S, Pimenta M, Amandi A (2009) Personalized e-learning environments: considering students' contexts. In: Proceedings of WCCE 2009, world conference on computers in education IFIP—international federation for information processing, pp 48–57

