Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 8 (2017) pp. 2367-2380

http://www.ripublication.com

Intelligent Recommendation System using Semantic information for Web Information Retrieval

Anupama Prasanth

Lecturer, College of Computer Studies,

AMA International University , Salmabad, Bahrain.

Abstract

Due to increase in current websites, the web users have been offered with numerous choices and because of this there exists an inability in decision making by the web user while surfing the web. Thus, the user needs the effective and useful suggestion or recommendation for accessing the website efficiently. Therefore, the recommendation results are very useful for the user by handling the website in an efficient way. The core technique of the webpage recommendation rules is prediction and learning process. These processes are used for appraising what users would like to view in the future website and learn the users' behaviors. In particular, this predicting and learning process can suggest an interesting item from the huge set of item based on the knowledge obtained about an active user.

The proposed system predicts user navigational preferences from their previous activities and those learned pattern is used for recommending more preferable web sites to web users. Before recommending the pages to the user a collaborative filtering mechanism has been used to refer the historic visiting preferences of a user with other users of having the same interest.

Keywords: Recommendation System, Collaborative Filtering, Pattern generation, Web Information Retrieval, Semantic Analysis**.**

1. INTORDUCTION

In the WWW perspective, recommender frameworks are becoming widely accepted and utilized by web users and web information retrieval framework to perform results

of both recommendation and pre-fetching. Most researchers focus on WUM that examines web log files with a procedure of discovering knowledge in vast databases. Actually, the websites are creating a huge amount of web logs data that comprises help data about the behavior of the web user. The WUM termed as analysis and discovery of helpful information from the WWW. It is also considered as "the application of data mining methods to huge web data repositories" which is known as recognizing the access pattern of web user from different types of web servers in automatic manner. To facilitate the user web page access, the web recommendation system is required. The WUM process can be employed to predict the subsequent web page access.

The proposed system predicts user navigational preferences from their previous activities and those learned pattern is used for recommending more preferable web sites to web users. Before recommending the pages to the user a collaborative filtering mechanism has been used to refer the historic visiting preferences of a user with other users of having the same interest.

Thus, this paper gives an overview in recommendation systems and gimps on recommendation rule and semantic recommendation systems. Then its shows the proposed system. Further, this paper intended with the comparison results of these algorithms with the experimental work by using web log data. The performance results show that the proposed system CHAMELEON clustering with Semantic analysis combined with the Recommendation Rules Generation component performs better than the other two techniques.

2. RECOMMENDATION SYSTEM: AN OVERVIEW

The consequent inevitable and dramatic growth emergence of recommender systems such as recommending music or TV programs, movies, products that a user would identifying the interesting web pages or suggesting the way of searching for extracting needed information from the web server [6], [8]. The recommendation system enables personalization advantages of the web user. This recommendation system can be utilized to either predict where a specific user will like specific web page or to recognize a set of N items that would be of certain user interest which is named as Top N recommendation [7].

Recommendation systems are widely categories into two classes such as collaborative based filtering and content based filtering. In the content based filtering process each web user is independently accessed the particular websites.[1] In the collaborative filtering system, document based recommendation is made which is obtained from prior readers of the same document, thereby recognizing user communities with shared interest as shown in Figure 1

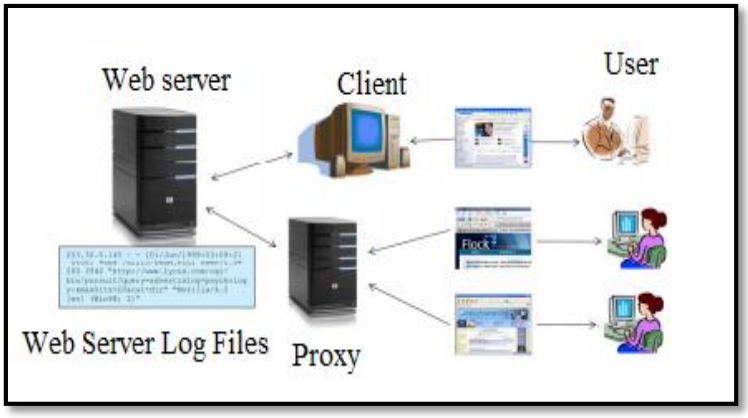

Figure 1 Users Browsing Activities

2.1 Recommendation Rules

The recommendation rules generation can access, in real time, the rules generation information taken over by pervious modules. Therefore, it will link to the session identification module and will consider from the single users' sessions, so as to generate recommendation rules. The generation of recommendation rule is being thru online in a transactional procedure [5]. Because of the innovating method of storing of recommendation rules and processing sessions by using semantically similar clustering process is succeeded in attaining a scalable and reliable recommendation model, able to work in the real time application with huge volume of data.

The component of recommendation rules generation system is utilized for searching the best access web page in the pattern as recommended by the web user's present sequence access. The applicability of recommendation rule generation system is to enhance by utilizing user's current access suffix sequences [12]. It will be measured when the corresponding access sequence's matching pattern is not found. An added advantage of proposed work is that the final set of recommendation can be ordered using semantic description knowledge of web content. Here the concluding set of recommendations might be in one or other sub classes not yet visited web documents or with associates the related web documents in the semantic document clusters.

The semantically similar pattern based recommendation engine is utilized for recommendation rules creation procedure and the web document clusters information has been fed into the semantic based recommendation engine as an input [11]. At this time already allocated corresponding cluster document where used for creating recommendation rules and corresponding URLs. The document which is matched to the user request, that web document recommends the corresponding URL to the web user.

2.2 Semantic Recommender System

The semantic recommender systems shows that the positive view point of web development. The incorporation of web mining with semantic knowledge plays an significant role in the robust recommender systems development [3]. Specifically, domain knowledge could be used to improve the web personalization procedure with the content mined from the documents, so that the documents can be clustered depends on the semantically similar measures. Then, web usage data is produce semantically improved navigational patterns. Afterwards, the system can build recommendations, relied upon the input patterns semantically associated with the created navigation patterns.

Using semantic analysis, the relationships between document classes handled by the web users can properly defined so as to create effective web mining models. The pattern mining can be obtained from the accessed web document to signify the web user needs. In such a case, the semantically similar cluster is helpful to search right data when a user creates a request. Additionally, the domain knowledge labeling the web user so as to create semantically enriches recommendation and also enhances the accuracy of prediction. Furthermore, a semantic recommender system [2] learns the semantic patterns (intersects) from the web log files, which provides recommendation.

In this work, the basic recommendation methods containing the content based, demographic, general-like and collaborative filtering are utilized to extract the preferences of the web user. Depends on the preferences, the interesting pattern can be selected [10]. Moreover, the semantically similar clustering mining process is used to produce the final user recommendation list. In that, the semantically similar clusters are used to define the web user preferences and the page for recommendation. It is created to be generalist so that the generated system can be handled with any kind of domain. The proposed system can fruitfully employed with different kinds of huge data sets.

3. PROPOSED RECOMMENDATION SYSTEM

The proposed work introduced a measure, “Page Popularity”, to count the number of times a page been accessed by a user during a certain time period. The entire period which considered for analysis is divided into two parts, “present” and “past” and the accesses in the recent have given more weights than history. The emerging topics and user latest interest can be easily identified by this way of analysis. This study has used access details of users from web server to analyze the usage and interest of web document over time.

Cluster web pages and identify the pages which are semantically similar to the user query during the time period. Chameleon hierarchical cluster and semantic analysis

processes are done through parallelly, the recommendation rules is employed in semantically similar data as shown in **Error! Reference source not found.**

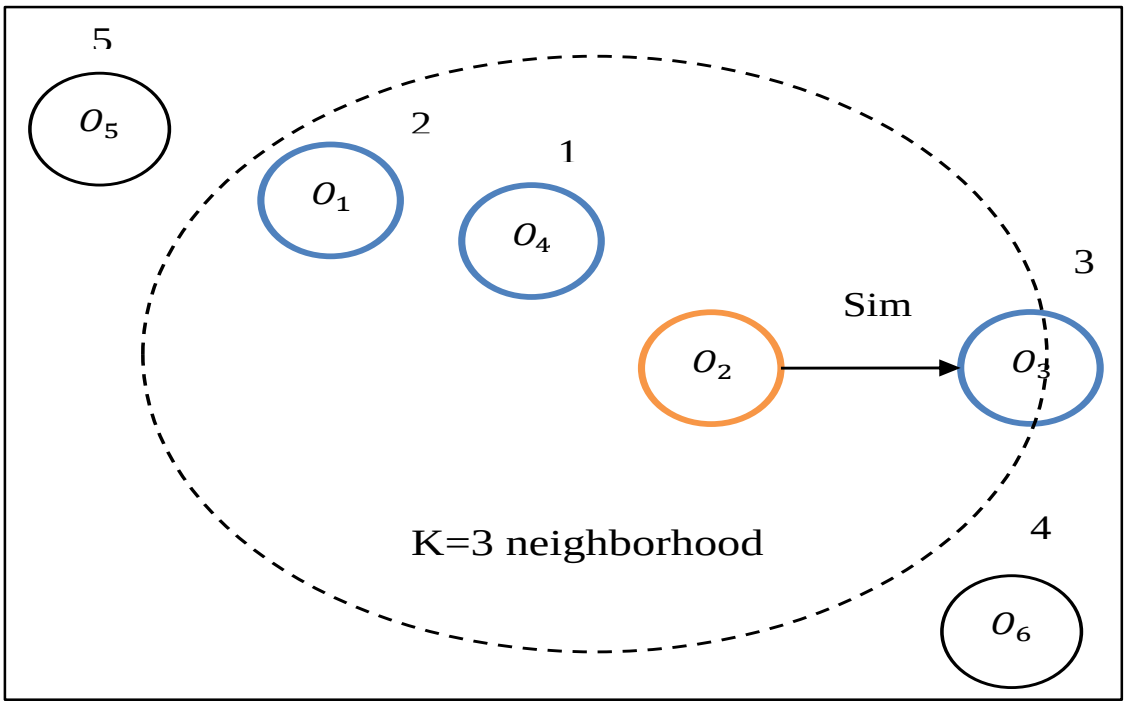

Figure 2 Semantically Clustering

The terms of similarity cluster objects, either by considering a given number of most similar objects (KNN) or all web users within a specified similarity threshold.

Relative Inter-Connectivity (RI) and Relative Closeness (RC), these two metrics are used for compute the similarity between two clusters point and these two metrics makes the absolute inter-connectivity between two different clusters. The RI is as follows

$$\mathrm{RI}\ (C_i C_j\ ) = \frac{\left|\mathrm{EC}_{\{C_i C_j\}}\right|}{\frac{\left|\mathrm{EC}_{C_i}+\mathrm{EC}_{C_j}\right|}{2}}$$

Where Edge-Cut (EC), is known as the edge that link C_i to C_j. The RC is as follows

$$\mathrm{RC}\ (C_i C_j\ ) = \frac{\bar{S}\mathrm{EC}(C_i,C_j)}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}\mathrm{EC}(C_i)+\frac{|C_i|}{|C_i|+|C_j|}\bar{S}\mathrm{EC}(C_j)}$$

Parallel with clustering, the domain knowledge extraction of web pages also done using appropriate tools. After that combine the domain knowledge with clustered pages, and construct vector space matrix with user query. Then calculate similarity measure using cosine similarity. This will provide the most semantically similar web pages of user query.

3.1 Page Popularity and Recommendation.

The page popularity is the measure used in this proposed work to measure the popularity of a page during a time period. This measure counts number of times a particular page is accessed in the given time duration. The frequency of usage gives the interest of the user on a page. The interest of a user on certain thing is highly depending on time. User interest may vary when time goes. So of things on which user showed much interest on certain time will not be having the same interest at present. Because of the dynamic nature of web page, the content and structure of the page also changes when time goes. The derivation of the measure page popularity is based on these facts, varying interest and dynamic web page. The varying interest of the user on a page is modeled by separating the time duration for the study into two, “past” and “present”. Then the pages accessed in recent have given more weightage than history. So the most popular pages in the recent period will get more weightage than the page which was popular earlier, and at the same time the page which was popular in the history is also having weightage so that also will be accounted for recommendation. As time changes, the substance, structure and use of a Web page changes. These progressions can be displayed both a solitary page level or for an accumulation of pages. Looking from a perspective of a solitary page, the idea that a Web page speaks to might change or advance concerning the time. Additionally, the fundamental structure of a page might change, i.e. the quantity of inlinks and the quantity of out connections might change. Subsequent to most auxiliary mining work considers that "if a page is indicated by some other page, then it supports the perspective of that page". So as the quantity of approaching connections changes, the theme that the page speaks to might change with timeframe. Likewise the adjustment in the quantity of out connections might mirror the adjustment in the pertinence of the page as for a specific theme. The utilization information is likewise influenced by the substance and basic change in a Web page. The utilization information realizes in data the theme the page is prevalent for. Also, this ubiquity could possibly be essentially reflected by the adjustment in the substance of the page or the pages indicating it. A page's notoriety might possibly be influenced by the adjustment in its indegree or outdegree. The Figure 3 depicts the basic concepts of popularity.

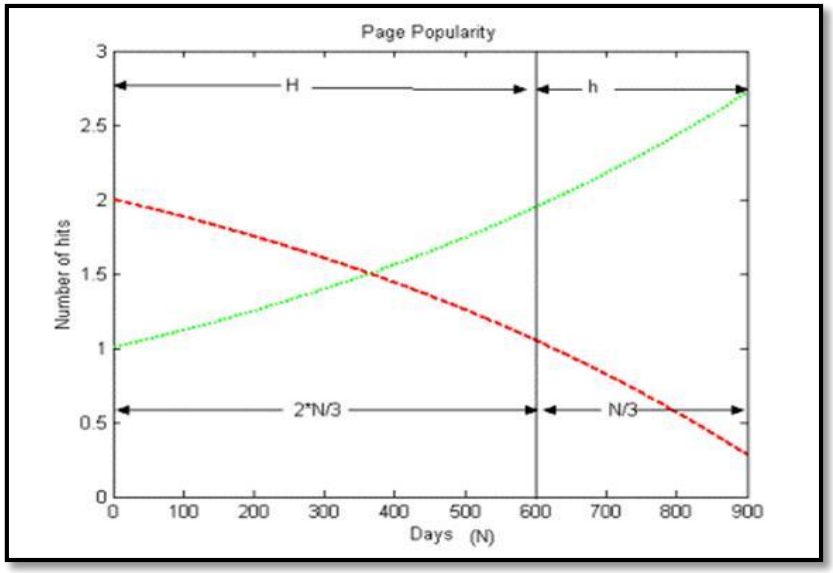

Figure 3 Concept of Page Popularity

The "page popularity" of each page is calculated based on its usage in the history. From the usage history, separate the usage into two parts, present history and past history. This sort of a thing should be possible by simply considering a "present" time of data that would bring about loss of information of the old data. The ideal solution for no loss of data is to consider longer duration and give more weightage to the present history than past, so the loss of data will be less. The idea in the **Error! Reference source not found.** is that the red and green shows the number of hits of certain pages during the time interval N. In the past history, the page which represents red has higher access rate, but in the present history, it redefined by the page represented by green. Though the page having red was popular in history, but at present the page green is more popular hence more weightage, so that can be considered as a newer topic or something that is picking up fame rather than the web page that is spoken to by the red curve, which is no more utilized that much.

So the page popularity

$$PagePopularity(w_i) = \frac{1}{3} \cdot \frac{H}{(2N/3)} + \frac{2}{3} * \frac{h}{N/3} = \left(\frac{1}{2N}\right) * \frac{(H + 4 \cdot h)}{H + h}$$

Here N is the total number of days, duration, considered for analysis. H and h are the number of hits in past and recent. This is really the normal number of hits amid the "past" time period and the "present" time period. Normal was considered as it would kill the impact of any sudden spikes or drops in use every day. On the off chance that we weigh as indicated by some other scale such as direct, such sudden changes might radically help or cut down the rank of a page. To give significance in the late

utilization, considered the initial two-third of the time as "past" history and last one-third as "present" history. There was no specific motivation to pick thus, however it appeared a sensible evaluation. We then measured the hits in the "present" history twice as that of the hits "previously" history.

After calculating the page popularity of each page which is semantically similar to the user query, calculate their recommendation score using the formula:

$$RS(p_i)=\sqrt{w_i^p X\ sim(s_q,s_p)}$$

Rank the pages based on their recommendation score. In general there are different types of factors are consider for defining the recommendation set and these kind of factors include following activities

The matching process for each and every cluster based on their similarity to the user's present active session.

Whether, the active URLs for recommendation have been activated by the current user in the present active session.

Short-term history deepness of the present web user is signifying the user's activity history portion. This portion is considered as relevant for the recommendation.

The lengthy physical link path from the present user's active session is characterized as candidate URL and is used for recommendation.

In this study, additionally includes the idea of recommendation is depends of avoiding the set of mismatch patterns and it can improves the entire processes. To avoiding the mismatch patterns are initially created by comparing the process models, specifically which are processing better way in terms of considering similarity value.

4. RESULT AND DISCUSSION

This study has taken publically available web log data on the internet as the usage data for experiments. The data available on internet is either unprocessed data or in processed format. The data set used for this study has taken from https://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data.
Usually the knowledge discovery methods for data mining purpose have been using commonly-used data source for testing and comparing purposes.

The log was gathered from January, 2015 through May, 2015, a total of 5 months. There are 172358 entries user accesses were registered. The values from the dataset are modified for confidentiality reasons.

There are four main metrics are used for evaluating the quality and efficiency of the semantically cluster recommendation system such as precision, coverage, F1-Measure and R-Measure.

The testing is executed for 150 different kinds of web pages in the proposed framework. The length of user transaction history (web log) is set at 153 days. The proposed system is tested for 50 such web users and the parameters are constant metrics. The following metrics and graphs explain the proposed system results.

In this study, we have evaluated the quality of the system with three similar types of algorithms. They are Modified Naïve Bayesian Classification with Supervise Learning Technique, Top N weighted Scoring Scheme and WebPUM. The results shows that the proposed algorithm is having a promising result, when compared with the existing popular algorithm in terms of Precision, Recall, F Measure and response time.

By precision it means how accurately a specific framework functions. Precision is defined as the extent of records retrieved that is important. In precision the non-relevant items is discarded by the user. The general formula for calculation of precision may be state as:

$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}}$$

Error! Reference source not found. shows the precision results in term of different cluster size. The result clearly shows that the precision rate of proposed system is better than other three similar algorithms. Even with varying cluster size, the proposed system Semantic Chameleon Clustering Algorithm with Recommendation rule shows more weightage than others.

The precision evaluation value is high when there is relevant information recommendation. Thus, a proposed semantically clustered recommendation framework has a high accuracy than other existing methods. The precision evaluation performance result is shown in **Error! Reference source not found.** describes that the proposed method shows the promising result.

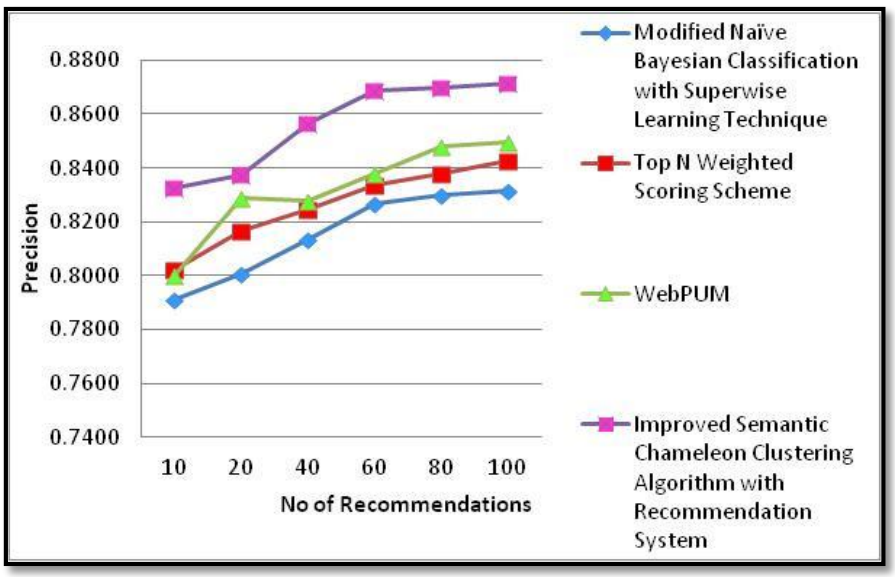

Figure 4: Precision

The term recall alludes to a measure of whether a specific thing is retrieved or the degree to which the retrieval of needed things happens. Whenever a user puts his/her query, it is the responsibility if the system to retrieve each one of those things that is applicable to the given query. When the collection is large it is not possible to retrieve all the relevant items. Thus, a system is ready to retrieve an extent of the aggregate significant report in response to a given query. The execution of a framework is regularly measured by review proportion, which indicates the rates of applicable things retrieved in a given circumstance.

The general formula for calculation of recall may be state as: Number of relevant item retrieved

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items in the collection}}$$

The recall evaluation value is high when there is relevant information recommendation. Thus, a proposed semantically clustered recommendation framework has a high accuracy than other existing methods. The recall evaluation performance result is shown in Figure 5 describes that the proposed method shows the promising result.

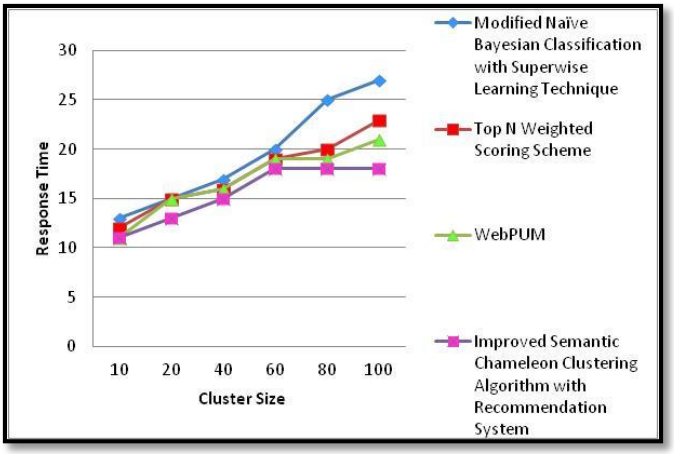

Figure 5 : Recall

The F measure is based on precision and recall value. It is one of the most popular measures used for measuring cluster accuracy.

$$F(i,j))=\frac{2*\text{precision }(i,j)*\text{recall}(i,j)}{\text{precision }(i,j)+\text{recall}(i,j)}$$

The F-Measure evaluation value is high when there is relevant information recommendation. Thus, a proposed semantically clustered recommendation framework has a high accuracy than other existing methods. The F-Measure evaluation performance result is shown in **Error! Reference source not found.** describes that the proposed method shows the promising result.

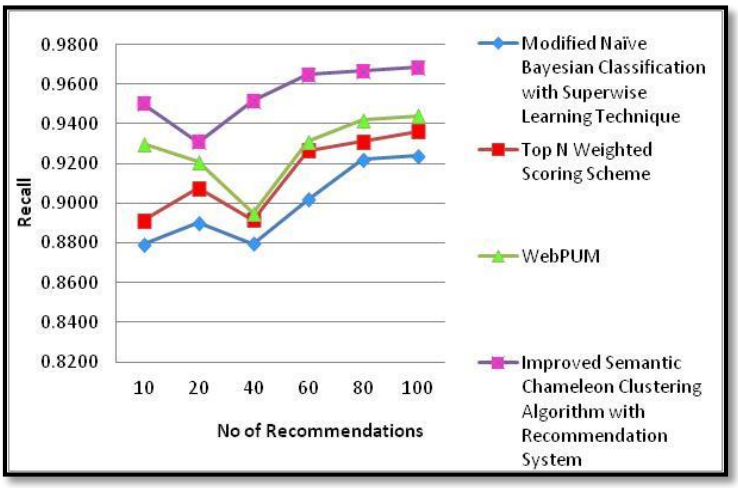

Figure 5: F – Measure

One of the major challenges of all IR systems is to provide information to the user as immediately as possible. So the metric response time is very important to measure the efficiency of IR system. Response time is nothing but, which is the time gap between users gives a query as the beginning of input and the end of the response. The experiment on response time of the system based on query "movie news" conducted on the same data set also proved that the proposed system has better response time than other systems.

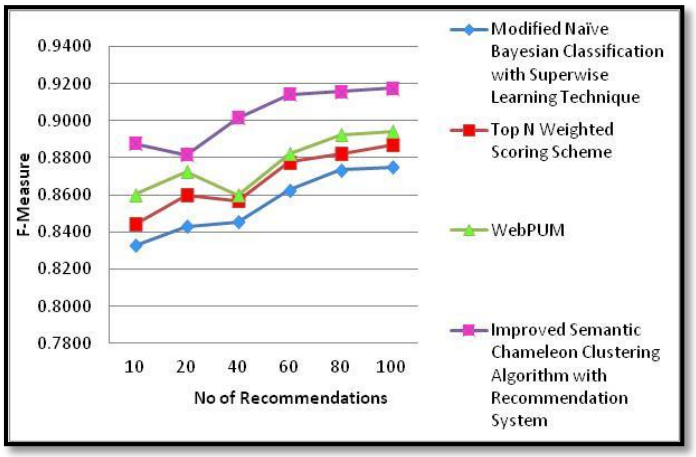

Figure 6: Response Time

From the **Error! Reference source not found.**, it is clear that as the cluster size increases, the proposed system need less time than other algorithms in responding to the query.

The response time comparison shows that the proposed system respond faster than other existing systems.

5. CONCLUSION

In this chapter present the web recommendation process depends on the web user's current web log history. To applying the frequent pattern mining process is used for strong recommendation system which helps the present searching process. These processes are done by using semantic relation of the cluster data. The Chameleon hierarchical clustering process is used for discovering the frequent search pattern. The semantic analysis is employed for creating frequent search pattern and this process makes the semantically similar patterns. The web user's profiles are clustered with the intention of alternative searching process which makes use the entire user log files. Hence this procedure is minimizing the searching time and greatly improving the performance of the overall recommendation process.

REFERENCES

[1] Robin Burke, "Hybrid Recommender System: Survey and Experiments ", ACM Communication, 2002 .377-408.

[2] Gemmis, M. D., Lops, P., Semeraro, G., & Basile, P. "Integrating Tags in a Semantic Content based Recommender", Proceedings of ACM Conference on Recommender Systems, 163-170.2008.

[3] Houda, O., Omar, N., & Philippe, B. . Minimum Redundancy and Maximum Relevance for Single and Multi-document Arabic Text Summarization. Science Direct 2010.

[4] Jafari, M., Sabzchi, F. S., & Irani, A. J.. Applying Web Usage Mining Techniques to Design Effective web Recommendation Systems: A Case Study. Advances in Computer Science: An International Journal , 78 – 90, 2014.

[5] Jajvand, A., Jeyyadi, M. A., & Salajeghesh, A. A Hybrid Recommender System for Service Discovery. International Journal of Innovative Research in Computer and Communication Engineering , 1342 – 1347, 2013.

[6] Jalali, M., Mustapha, N., S, M. N., & Mamat, A. (2010). WebPUM: A Web-based Recoomendation System to Predict User Future Movements. Elseiver - Expert Systems with Applications , 6201 – 6212, 2010.

[7] Murtadha, Y. A., Sulaiman, M. N., Mustapha, M., & Udzir, N. I. (2011). Improved Web Page Recomendation Systems based on web Usage Mining. 3rd International Conference on Computing and Informatics (pp. 32 - 36). Bandung, Indonesia: ICOCI, 2011.

[8] Y.S, S., Mahadevan, D. G., & Prakash, M. (2011). Recommendation System Based on Web Usage Mining and Semantic Web : A Survey. International Journal on Recent Trends in Engineering and Technology , 97 – 100, 2011.

[9] T.Vijaya Kumar, H.S.Guruprasad, "Clustering Of Web Usage Data Using Chameleon Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 6, 2014.

[10] Suresh Shirgave, Prakash Kulkarni, "Semantically Enriched Web Usage Mining For Predicting User Future Movements", International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.4, 2013.

[11] Soheila Abrishami, Mahmoud Naghibzadeh, Mehrdad Jalali, "Web Page Recommendation Based on Semantic Web Usage Mining", Social Informatics

Lecture Notes in Computer Science, Springer, Vol.7710, pp 393-405, 2012.

[12] C.Ramesh, K. V. Chalapati Rao, A.Goverdhan, "A Semantically Enriched Web Usage Based Recommendation Model", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 5, 2011.