An Enhancement of Svm Based Semantically Enriched Variable Length Confidence Pruned Markov Chain Model Based Web Page Recommendation System

R. Rooba

***Abstract** Semantic Variable Length Markov Chain Length Model (SVLMC) is a web page recommendation system which combined the fields of semantic web and web usage mining by the Markov transition probability matrix with rich semantic information extracted from web pages. Though it has high prediction accuracy, it has problem of high state space complexity. The high space complexity reduce the execution speed and reduce the performance of the system, which was resolved by Semantic Variable Length confidence pruned Markov Chain Model (SVLCPMC) model that provides high user satisfied recommendation and Confidence Pruned Markov Model (CPMM). The time consumption of CPMM was reduced by Support Vector Machine (SVM). But still the recommendation accuracy is still below the user satisfaction. So in this paper, quickest change detection using Kullback-Leibler Divergence method is introduced to improve the accuracy of recommendation generation by developing a scalable quickest change detection schemes that can be implemented recursively in a more complicated scenario of Markov model and it is included in the training data of SVM. Then the performance of web page recommendation is improved by ranking the web pages using page ranking technique. Thus the performance of web page recommendation generation system has been improved. The experiments are conducted to prove the effectiveness of the proposed work in terms of prediction accuracy, precision, recall, F1-measure, coverage and R measure.*

***Keywords** : Web Page Recommendation, Semantic Variable Length Markov Chain Length Model, quickest change detection, Kullback-Leibler Divergence, Page ranking.*

I. INTRODUCTION

Prediction of user web navigation behavior [1] is becoming a tedious process because of the growth of the World Wide Web is increasing rapidly.

Manuscript published on January 30, 2020.
* Correspondence Author

Dr. R. Rooba*, Assistant Professor, Department of Computer Technology and Information Technology, Kongu Arts and Science College (Autonomous), Erode, Tamilnadu, India. E-mail: rrooba@gmail.com

Web usage mining is one of the applications of data mining techniques that automatically discover and analysis the patterns in click stream.

The main aim of web usage mining is to model, capture and analyze the user’s behavioral patterns and user’s profiles interacting with a web site. The discovered patterns are signified as collection of resources, pages or objects which are repeatedly accessed by a group of users with common interest or needs. It is also called as web log mining. Web log mining is a technique of interpreting and discovering of patterns of users accessing the web by mining the web log data. Web usage mining techniques are mainly based on association rules, clustering techniques, sequential patterns and Markov Models. Association rules create rules through the data mining methods to detect the user interesting patterns. While generating rules for huge volume of dataset this technique takes too much of time to predict the users interesting patterns. Whereas sequential pattern mining identified all patterns of sequence with the help of user specified minimum support. But it is more difficult to predict the navigation of pattern in sub sequences.

Markov models were more popularly utilized in the prediction of user’s next link of choice and it also predicts longer sequence of user’s navigation patterns and for pre-fetching links. The higher order of predicting models has major issues like high state space complexity, reduced coverage and sometimes reduced the prediction accuracy. Some of these problems was overcome by training the varying order of Markov models and utilized them in the prediction phase. This approach overcomes only the problem of prediction accuracy and state complexity. Semantic Variable Length Markov Chain Model (SVLMC) [2] was used to improve the recommendation accuracy by combining detailed semantic data with recommendation process. However, SVLMC model doesn’t consider the out link of the state that also influences the prediction accuracy. In order to enhance the prediction accuracy and remove the low accuracy states in the model Confidence-Pruned Markov Model (CPMM) [3] was used with SVLMC. CPMM consider both out-links and in-links of the state during pruning process. It improves the prediction accuracy by using a cloning concept.

The SVM was introduced to update the transition probability of new session link. However, the recommendation accuracy is still below the user satisfaction. In this paper, highly accurate recommendation system is obtained by introducing Kullback-Leibler Divergence [4] to develop scalable quickest change detection[5] schemes that can be implemented recursively in a more complicated scenario of Markov model. Though this approach improves the accuracy of recommendations, still better performance can be achieved by ranking the web pages.

Here PageRank is used to rank the web pages based on the in links and outlinks in a web page. The PageRank technique is modified by making link matrix as a column stochastic matrix. The performance of PageRank technique is also enhanced by introducing adaptive weighted gradient estimation to rank the web pages. Hence the highly accurate recommendation is achieved by Kullback-Leibler Divergence and adaptive weighted gradient estimation [6].

II. LITERATURE SURVEY

A semantic web mining approach [7] was proposed to find out the periodic web access patterns. It can be achieved by using web usage logs that integrates the information like consumer behavior and emotions through behavioral tracking and self reporting. Additionally fuzzy logic was used to signify requested resource attributes and real life temporal concepts of periodic pattern based web access activities. Thus these both representations consist of emotional and behavioral cues that were integrated with a Personal Web Usage Lattice which models the web access activities. By using created a Personal Web Usage Ontology that provided semantic web applications such as personalized web resources recommendation. However, different durations of periodic conditions and different number of personalized resources affect would affect the results in the absence of emotional influences.

A recommendation system was proposed namely WebPUM [8] to predict the user's future request of a particular website. The WebPUM was an online prediction that utilized a web usage mining system. Additionally a novel approach was proposed that classified a user navigation pattern which was used to predict the user's future intension. The novel approach models user navigation pattern with the help of new graph partitioning algorithm. Then these patterns were used in navigation pattern mining phase. The navigation patterns were classified based on common subsequence algorithm. However, the recommendation is still below the user satisfaction.

A web navigation framework was proposed called Web Navigation Prediction Framework for Web page Recommendation (WNPWR) [9] to predict the user's future request. The presented framework created and generated a classifier based on sessions. The sessions were utilized as training samples for prediction of web page recommendation. By computing the average time on visiting pages the training samples were created and classifiers were generated by the N th Markov Models. Each session were mapped the created classifiers. Where there is any session mapped to more than one classifier then only one classifier was used to map the each session with the help of pagerank algorithm in filtering process. The filtered data were used to train Support Vector Machine. Then the trained SVM was used for web page recommendation. The major disadvantage of this work is training and testing speed of SVM is high.

A technique and a framework was developed [10] for frequent web navigation patterns. In the pattern generation process, semantic information plays an important role in the web usage mining techniques. The presented technique and framework combined semantic information with the web navigation patterns. In this frequent navigational patterns were created by ontology instances other than the web page address. Thus the web page navigational patterns reflect the semantics of semantics of navigational behavior more accurately and explicitly. However, the presented technique has high time consumption.

The user web navigation session was modeled by a proposed [11] called Dynamic Nested Markov Model that structured the Markov model. This presented model utilized the nesting concept where the higher order Markov model was nested with the lower order Markov Model. By nesting the higher order model into the lower order model the second order Markov model was presented inside the first order Markov Model. The benefits of both the lower order model and higher order model were obtained through the Dynamic Nested Markov Model. However, the large number of states in the model degrades the performance of prediction of next web page accessed by the users.

An approach was proposed [12] to analysis the navigational behavior of user. The proposed approach used Grey Relation Pattern Analysis (GRPA) with variable length Markov Chain. It considered the sequential information in web usage data and this was extended by combining the GPRA with variable length Markov Chain with a web user navigation behavior model that improves the web usage mining applications. The variable-length Markov chain (VLMC) models offered the probability of the next link selected when visiting a Web page while taking into consideration the trail followed to reach that page. Thus this proposed approach works better than the clustering methods.

A novel method was proposed [13] for better web page recommendation which was achieved through integrating the web usage and domain knowledge of a web site. The domain knowledge was represented by using proposed two models. One of the proposed models utilized ontology to signify the domain knowledge and another proposed model utilized one automatically created semantic network to signified web pages, domain terms and the relations among them. Additionally a new model was proposed called conceptual prediction model that automatically created semantic networks of the semantic web usage knowledge which was the combination of web usage knowledge and domain knowledge. A web page candidate was generated based on the creation of a number of effective queries that represents a set of recommendation. However, these proposed methods were tested only with the recommendation length of 5.

III. METHODOLOGY

In this section, the quickest change of transition probability between old session and new session are detected using Kullback-Leibler Divergence method. From this process the quickest change between the new session and old session which is more useful for web page recommendation.

The quickest change detection using Kullback-Leibler Divergence method is used in the training data of SVM. Then page ranking techniques like PageRank, modified PageRank and Adaptive Weighted Estimation PageRank is used to rank the web pages based on the similarity score matrix which also plays an important role in web page recommendation.

A. Quickest Change Detection

Initially the web pages are pre-processed by web pre-processing components that observed the user web log files and determine the web navigation session of user. Then the next link choice is predicted using semantically enriched matrix W that generates the recommendations. The transition probability and similarity score matrix are considered to generate the recommendations. The semantically enriched matrix is given as follows

$$w_{p_i p_j} = T_{p_i p_j} + \begin{cases} (1-\alpha) * M_{p_i p_j}, M_{p_i p_j} & > 0, \\ 0 \quad , \quad M_{p_i p_j} & = 0, \end{cases}$$

-----------(1)

In the above equation 1, $T_{p_i p_j}$ denotes the transition probability matrix determined Semantic Variable Length confidence pruned Markov Chain Model (SSVLCPMC) model [11] and $M_{p_i p_j}$ denotes the similarity score matrix and α denotes the semantic coefficient factor. Based on weighted matrix the next link choice is created and the recommendations are generated.

The following table 1 shows the semantic enriched SSVLCPMC matrix.

Table - I. Semantic enriched SSVLCPMC matrix

	0	1	2	3	4	5	6	7	8
0	0.0	0.42	0.57	0.00	0.00	0.00	0.00	0.00	0.00
1	0.0	0.00	0.45	0.12	0.00	0.00	0.00	0.00	0.00
2	0.0	0.00	0.00	0.6	0.40	0.00	0.00	0.00	0.00
3	0.0	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.12
4	0.0	0.00	0.00	0.00	0.00	0.33	0.00	0.27	0.40
5	0.0	0.00	0.00	0.00	0.00	0.00	1	0.29	0.00
6	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38
7	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1
8	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

In the web page recommendation, it is assumed that at some unknown time, some new sessions occurs in the web page navigation and changes the distribution of observations abruptly from their normal patterns. So the main aim is to detect the true change as soon as possible under a restriction on the false alarms. In this paper, the Kullback-Leibler Divergence method is used for quickest change detection. The quickest change detection using Kullback-Leibler Divergence method is included in the training data of SVM. Kullback-Leibler Divergence detects a change on the transition probability between old session s_0 and new session s_1. For any transition probability Γ satisfying false alarm constraints, it is explained below

$$D(\Gamma) \geq (1 + o(1)) \frac{\log\gamma}{K(s_0, s_1)} \qquad (2)$$

In the above equation 2, $\gamma \to \infty$, $K(s_0, s_1)$ is the Kullback-Leibler Divergence is defined as follows:

$$K(s_0,s_1) = \lim_{n\to\infty} \frac{1}{n} \log s_n \qquad (3)$$

In the above equation s_n denotes the likelihood ratio statistic for first n sessions $(s_1, \dots, s_n)$. A naive approach is the Monte Carlo simulation method that estimates $K(s_0, s_1)$ by $\frac{1}{n} \log s_n$ for large n under probability P_{s_1}. The log likelihhod ratio statistic $\log s_n$ can be computed as:

$$\log s_n = \sum_{i=1}^{n} \log(p_{s_i}) - \sum_{i=1}^{n} \log(p_{s_i+1}) \qquad (4)$$

In above equation 4, the log likelihood ratio statistics between n sessions are computed. The quickest change detection using Kullback-Leibler Divergence is included with the training data of SVM. Then the weighted matrix becomes

$$w_{p_i p_j} = D(T_{p_i p_j}) + \begin{cases} (1-\alpha) * M_{p_i p_j}, M_{p_i p_j} & > 0, \\ 0 \quad , \quad M_{p_i p_j} & = 0, \end{cases} \qquad (5)$$

In above equation 5, the weighted matrix is constructed based on the quickest change detection of transition probability, similarity score matrix and the semantic coefficient factor.

The weighted matrix based on calculation of Kullback-Leibler Divergence is given as follows:

$$\log s_n = \sum_{i=1}^{n} \log(p_{s_i}) - \sum_{i=1}^{n} \log(p_{s_i+1})$$

For page 0 t0 1,

$$\log s_0 = \sum_{i=1}^{n} \log(p_{s_0}) - \sum_{i=1}^{n} \log(p_{s_1})$$

$\log s_0$=

$(\log 0.42 + \log 0.57) - (\log 0.45 + \log 0.12)$

= (-0.377-0.244)-(-0.347-0.921)

= -0.621+0.574

= -0.047

$K(s_0, s_1) = \lim_{n\to\infty} \frac{1}{n} \log s_n$

$K(s_0, s_1) = \frac{1}{10} \times 0.047 = 0.0047$

For page 0 to 2,

$\log s_0$=$(\log 0.42 + \log 0.57) - (\log 0.6 + \log 0.4)$

=-0.621+0.62=-0.001

$K(s_0, s_2) = \frac{1}{10} \times 0.001 = 0.0001$

$K(s_0, s_2)$=0.0001

Similarly Kullback-Leibler Divergence is calculated for all pages. Then the weighted matrix is given as follows:

Table -II. Weighted matrix based on Kullback-Leibler Divergence

	0	1	2	3	4	5	6	7	8
0	0.00	0.0047	0.0001	0.1346	0.0827	0.0083	0.0201	0.0621	0.0621
1	0.0047	0.00	0.0046	0.0377	0.0874	0.0036	0.0154	0.0574	0.0574
2	0.0001	0.0046	0.00	0.0423	0.0828	0.0082	0.0199	0.062	0.62
3	0.0424	0.0376	0.0423	0.00	0.1251	0.0341	0.0223	0.0197	0.0197
4	0.0827	0.0874	0.0828	0.1251	0.00	0.091	0.1027	0.1448	0.1448
5	0.0083	0.0036	0.0082	0.0341	0.091	0.00	0.0118	0.0538	0.0538
6	0.0201	0.0154	0.0199	0.0223	0.1027	0.0118	0.00	0.0420	0.0420
7	0.0621	0.00574	0.062	0.0197	0.1448	0.0538	0.0420	0.00	0.00
8	0.0621	0.00574	0.062	0.0197	0.1448	0.0538	0.0420	0.00	0.00

B. Web Page Recommendation with Page Ranking

Ranking the web pages is done in terms page contents, structure, and usage statistics, helps in reducing the irrelevant pages thus improving the accuracy of recommendation. The PageRank algorithm is a key technique of Google to retrieve the more relevant web pages. Let us assume a web with x pages indexed by an integers y, $1 \leq j \leq x$ and $x \geq 2$. A directed graph is used to describe the web structure $G = (V, E)$ with node set $V = \{1,2\ldots,x\}$ and the edge set $E \subseteq V \times V$. If a web page m has a hyperlink pointing to a web page n, then $(m,n) \in E$. We call m the inlink of n and n the outbound of m. num_m denotes the set of all inlinks of m and num_n denotes the number of outlinks of n. The link matrix of the directed graph G is written as follows:

$$A = |a_{mn}| \in R^{x \times x}, a_{mn} = \begin{cases} \frac{1}{x_n}, if n \in num_m \\ 0, otherwise \end{cases} \qquad (6)$$

From the above equation 6, it is known that $\sum_{m=1}^{x} a_{mn}$ equals either 0 or 1 it is represented by $h_m^* \in [0,1]$ the importance score of the web page m and assume $\sum_{m=1}^{x} h_m^* = 1$. If $h_m^* > h_n^*$, then it is understand that the web page m is more important than the web page n. the importance score of each web page is calculated by number of inlinks and the corresponding importance scores.

$$h_m^* = \sum_{n \in num_m} \frac{h_n^*}{x_n} \qquad (7)$$

Then $h^* = (h_1^*, \ldots h_z^*)^T$ satisfies the following equation:

$$h^* = Ah^*, h^* \in S_p^x \qquad (8)$$

In the above equation 8, $S_p^x \triangleq h = \{h_1, \ldots, h_z\}^T R^x$ and $h_i \geq 0, i = 1, \ldots z, \sum_{i=1}^{z} h_i = 1$.

In order to ensure the uniqueness of h^*, the link matrix A is modified as all the web pages are artificially set to be outlinks of all dangling nodes. In other words, the 0_x columns of link matrix is replaced by $\frac{1_x}{x}$ that makes link matrix to be a column stochastic matrix $H = [h_{mn}] \in R^{x \times x}$ is described as $h_{mn} \geq 0, m, n - 1, \ldots x$ and $\sum_{m=1}^{z} h_{mn}$. The modified link matrix L is defined as follows:

$$L = (1 - \delta)H + \delta \frac{1}{x} 1_x 1_x^T \qquad (9)$$

In the above equation 9, δ denotes the damping factor ranges from 0 to 1. In the PageRank A and h^* with L and h_g^* respectively in equation 7, then the PageRank is defined as finding the vector $h_g^* = (h_{1g}^*, \ldots h_{xg}^*)$ which satisfies the following equation:

$$h_g^* = Lh_g^*, h^* \in S_p^x \qquad (10)$$

The pages are ranked by using following equations 11 and 12.

$$h(k+1) = Lh(k) = (1 - \delta)Hh(k) + \frac{\delta}{x} 1_x \qquad (11)$$

The stochastic based page ranking is improved by adaptive weighted gradient estimation based page ranking. The gradient estimation $\hat{g}_k(.)$ is defined as follows:

$$\hat{g}_k^*(\hat{\theta}_k) = \beta_1 \hat{g}_k(\hat{\theta}_k) + \beta_2 \hat{g}_{k-1}(\hat{\theta}_{k-1}) + \cdots + \beta_m \hat{g}_{k-m+1}(\hat{\theta}_{k-m+1}) \qquad (12)$$

Some assumptions are considered to rank the web pages which are defined as follows:

In the above equation 12, $\beta_1, \ldots \beta_m$ are positive numbers are positive numbers (the weights) that fall in [0, 1], which can determine how much impact each element would have on the final decision of the gradient estimation. $\hat{g}_k(\hat{\theta}_k), \hat{g}_{k-1}(\hat{\theta}_{k-1}), \ldots \hat{g}_{k-m+1}(\hat{\theta}_{k-m+1})$ are the gradient estimations at *m* previous time steps. The ranking condition is defined as:

$$h_k(\hat{\theta}_k) = \hat{g}_k(\hat{\theta}_k) - E\left(\hat{g}_k(\hat{\theta}_k)\middle|\hat{\theta}_k\right) \qquad (13)$$

where E(.) is the expected value.

The equation 13, can be rewritten as follows:

$$(\hat{\theta}_{k+1}) = (\hat{\theta}_k) - a_k[\hat{g}_k(\hat{\theta}_k)] + b_k \hat{g}_k(\hat{\theta}_k) + e_k(\hat{\theta}_k) \qquad (14)$$

where, $b_k = E\left(\hat{g}_k(\hat{\theta}_k) - \hat{g}_k(\hat{\theta}_k)\middle|\hat{\theta}_k\right), a_k$ is defined as gain sequence and e_k is the error term.

Based on equation 13, the page ranking equation 12 can be rewritten as

$$h(k+1) = Lh_k(\hat{\theta}_k) = (1 - \delta)Hh_k(\hat{\theta}_k) + \frac{\delta}{x} 1_x \qquad (15)$$

Let, $\quad h(k+1) = \begin{pmatrix} 0.5 \\ 0.62 \\ 0.43 \\ 0.87 \\ 0.21 \\ 0.54 \\ 0.32 \\ 0.75 \\ 0.21 \end{pmatrix}$

Table- III. Refined recommendation matrix by stochastic gradient page ranking

	0	1	2	3	4	5	6	7	8
0	0.00	0.00291	0.000043	0.1171	0.01737	0.00448	0.00643	0.04657	0.01304
1	0.00235	0.00	0.00198	0.03279	0.01835	0.00194	0.00493	0.04305	0.01205
2	0.00005	0.00285	0.00	0.03680	0.01739	0.00443	0.00637	0.0465	0.1302
3	0.0212	0.02331	0.01819	0.00	0.02627	0.01841	0.00714	0.01478	0.00414
4	0.04135	0.05419	0.03560	0.10884	0.00	0.04914	0.03287	0.1086	0.03041
5	0.00415	0.00223	0.00353	0.02967	0.01911	0.00	0.00378	0.04035	0.01129
6	0.01005	0.00955	0.00856	0.01940	0.02157	0.00637	0.00	0.0315	0.00882
7	0.03105	0.00356	0.02666	0.01714	0.03041	0.02901	0.01344	0.00	0.00
8	0.03105	0.00356	0.02666	0.01714	0.03041	0.02901	0.01344	0.00	0.00

IV. EXPERIMENTAL RESULTS

The proposed result is tested in DBpedia dataset from Semantic Web dog food Web site. A small number of requested pages removed by the filtering. The number of users calculated by based on access entries, clean access entries, accessed web pages in log. The web crawler is mainly responsible for extracting the page content in the web. The following four methods mainly used to differentiate the existing system and proposed system i.e., prediction accuracy, precision coverage, F1 measure and R measure. The F1 measure, precision and coverage are inversely related to each other so they provide equal weight.

A. Precision.

Let us assume that, a set of user sessions 't' from the test set is viewed as web pages and window of size $|w|$ produces recommendation set Rec, using recommendation engine. Precision is a number of relevant Web pages retrieved divided by the total number of Web pages in recommendations set. Thus precision of Rec with respect to t is given by,

$$Precision\ (Rec, t) = \frac{|\ Rec\ \cap (t - w)|}{|\ Rec\ |}$$

where $|\ Rec\ \cap (t - w)|$ - number of common Web pages in both recommendation set and evaluation set.

The Ssvlcpmc, Ssvlcpmc-Qcd, Ssvlcpmc-Qcd-Pr, Ssvlcpmc-Qcd-Mpr And Ssvlcpmc-Qcd-Awgepr model is compared. In figure 4.1, X axis will be taken as different models and Y axis will be taken as precision value. From the figure it is proved that the proposed SSVLCPMC-QCD-AWGEPR has high precision than the other models.

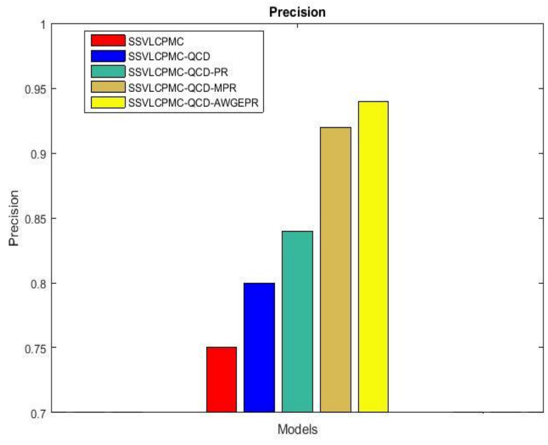

Fig. 4.1 Comparision of Precision

B. Coverage

Coverage is the ratio between the number of relevant Web pages retrieved and the total number of Web pages that actually belongs to the test user session. Coverage of Rec with respect to t is given by,

$$coverage\ (Rec, t) = \frac{|\ Rec\ \cap (t - w)|}{|t - w|}$$

Table – IV: Comparison of Coverage value of the proposed and Existing systems

Model Name	Precision Value
SSVLCPMC	0.75
SSVLCPMC-QCD	0.81
SSVLCPMC-QCD-QR	0.86
SSVLCPMC-QCD-MPR	0.92
SSVLCPMC-QCD-AWGEPR	0.95

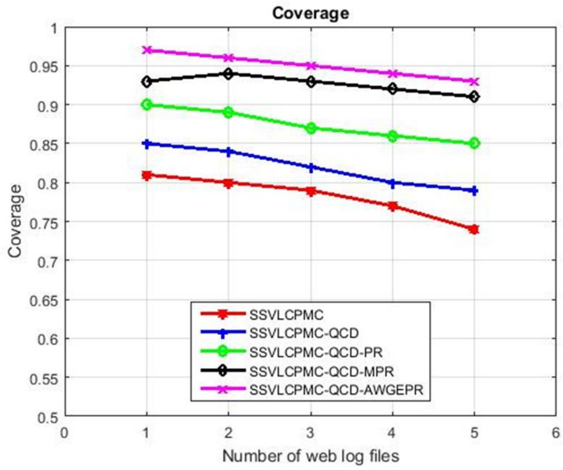

Fig. 4.2 Comparision of Coverage

The SSVLCPMC, SSVLCPMC-QCD, SSVLCPMC-QCD-PR, SSVLCPMC-QCD-MPR and SSVLCPMC-QCD-AWGEPR model is compared.

In figure 4.2, X axis will be taken as different number of web log files and Y axis will be taken as coverage value. From the figure it is proved that the proposed SSVLCPMC-QCD-AWGEPR has high coverage than the other models.

C. F1 – Measure

F1 measure is used to achieve high precision and high coverage. F1 measure is given by,

$$F1\ (Rec,t) = \frac{2\ \times precision\ (Rec,t) \times coverage\ (Rec,t)}{precision\ (Rec,t)\ +\ coverage\ (Rec,t)}$$

F1 measure achieves its maximum value when both precision and coverage are maximized.

The SSVLCPMC, SSVLCPMC-QCD, SSVLCPMC-QCD-PR, SSVLCPMC-QCD-MPR and SSVLCPMC-QCD-AWGEPR model is compared.

Table-V: Comparison of F1 – Measure value of the proposed and Existing systems

Model Name	F1 - Measure
SSVLCPMC	0.75
SSVLCPMC-QCD	0.81
SSVLCPMC-QCD-QR	0.86
SSVLCPMC-QCD-MPR	0.92
SSVLCPMC-QCD-AWGEPR	0.95

In figure 4.3, X axis will be taken as different models and Y axis will be taken as F1measure value. From the figure it is proved that the proposed SSVLCPMC-QCD-AWGEPR has high F1measure than the other models.

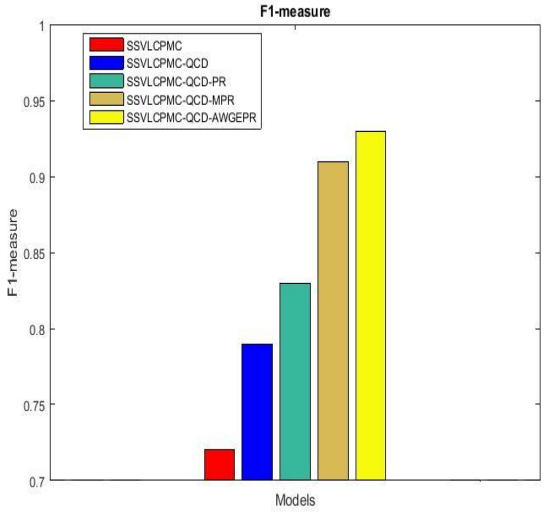

Fig. 4.3 Comparision of F1-measure

D. R – Measure

R measure is evaluated by dividing the coverage by the size of the recommendation set and it is given by,

$$R\ (Rec,t) = \frac{coverage\ (Rec,t)}{|\ Rec\ |}$$

Table-VI: Comparison of R– Measure value of the proposed and Existing systems

Model Name	R– Measure
SSVLCPMC	0.75
SSVLCPMC-QCD	0.81
SSVLCPMC-QCD-QR	0.86
SSVLCPMC-QCD-MPR	0.92
SSVLCPMC-QCD-AWGEPR	0.95

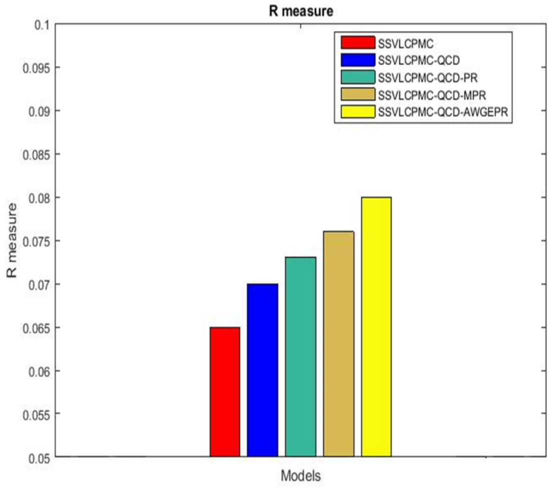

Fig. 4.4 Comparision of R-measure

The SSVLCPMC, SSVLCPMC-QCD, SSVLCPMC-QCD-PR, SSVLCPMC-QCD-MPR and SSVLCPMC-QCD-AWGEPR model is compared. In figure 4.4, X axis will be taken as different models and Y axis will be taken as R measure value. From the figure it is proved that the proposed SSVLCPMC-QCD-AWGEPR has high R measure than the other models.

E. Prediction Accuracy

Prediction accuracy is the measure of correctly recommended pages in all instances.

It can be calculated by

$$Accuracy = \frac{(True\ positive + True\ negative)}{(True\ positive + True\ negative + False\ positive + False\ negative)}$$

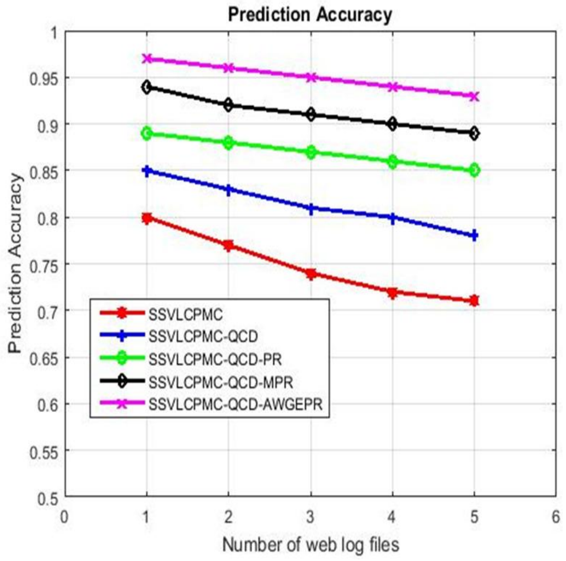

Fig. 4.5 Comparision of Prediction Accuracy

The SSVLCPMC, SSVLCPMC-QCD, SSVLCPMC - QCD- PR, SSVLCPMC –QCD -MPR and SSVLCPMC - QCD-AWGEPR model is compared. In figure 4.5, X axis will be taken as different number of web log files and Y axis will be taken as coverage value. From the figure it is proved that the proposed SSVLCPMC-QCD-AWGEPR has high coverage than the other models.

V. Conclusion

In this paper, the quickest change in transition probability between old session and new sessions are detected using Kullback-Leibler Divergence method. This is included in the training data of SVM. Then the performance of the web page recommendation is improved by using page ranking techniques. There are three different page rankibg techniques are used such as PageRank, modified PageRank and adaptive weighted gradient estimation page rank. The experiment results are conducted to prove the effectiveness of the proposed work in terms of prediction accuracy, f1-measure, precision, recall and R measure.

REFERENCES

1. Narvekar, M., & Banu, S. S. (2015). Predicting user's web navigation behavior using hybrid approach. *Procedia Computer Science*, *45*, 3-12.
2. Shirgave, S., Kulkarni, P., & Borges, J. (2014). Semantically Enriched Variable Length Markov Chain Model for Analysis of User Web Navigation Sessions. *International Journal of Information Technology & Decision Making*, *13*(04), 721-753.’
3. Popa, R., & Levendovszky, T. (2007, November). Marcov models for web access prediction. In *8th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics* (pp. 539-550).
4. J. R. Hershey and P. A. Olsen, “Approximating the Kullback Leibler divergence between Gaussian Mixture Models,” in Proc. of the International Conference on Audio, Speech and Signal Processing, Honolulu, Hawai, USA, April 15-20 2007, vol. 4, pp. IV–317.
5. Venugopal V. Veeravalli, Senior Member, IEEE, “Decentralized Quickest Change Detection”, IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 47, NO. 4, MAY 2001 (pp 1657 – 1663)
6. Jinlong Lei and Han-Fu Chen, *Fellow, IEEE, “Distributed Randomized PageRank Algorithm Basedon Stochastic Approximation”, IEEE TRANSACTIONS ON AUTOMATIC CONTROL, VOL. 60, NO. 6, JUNE 2015, pp(1641 -1646).*
7. Fong, A. C. M., Zhou, B., Hui, S., Tang, J., & Hong, G. (2012). Generation of personalized ontology based on consumer emotion and behavior analysis. *IEEE Transactions on affective computing*, *3*(2), 152-164.
8. Jalali, M., Mustapha, N., Sulaiman, M. N., & Mamat, A. (2010). WebPUM: A Web-based recommendation system to predict user future movements. *Expert Systems with Applications*, *37*(9), 6201-6212.
9. Sejal, D., Kamalakant, T., Tejaswi, V., Anvekar, D., Venugopal, K. R., Iyengar, S. S., & Patnaik, L. M. (2015, July). WNPWR: Web navigation prediction framework for webpage recommendation. In *Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on* (pp. 120-125). IEEE.
10. Senkul, P., & Salin, S. (2012). Improving pattern quality in web usage mining by using semantic information. *Knowledge and information systems*, *30*(3), 527-541.
11. Nigam, B., & Jain, S. (2010, November). Generating a new model for predicting the next accessed web page in web usage mining. In *Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on* (pp. 485-490). IEEE.
12. Madhuri, B. C., Chandulal, A. J., Ramya, K., & Phanidra, M. (2011). Analysis of users' web navigation behavior using grpa with variable length markov chains. *International Journal of Data Mining & Knowledge Management Process*, *1*(2).
13. Nguyen, T. T. S., Lu, H. Y., & Lu, J. (2014). Web-page recommendation based on web usage and domain knowledge. *IEEE Transactions on Knowledge and Data Engineering*, *26*(10), 2574-2587.

AUTHOR PROFILE

Dr. R.Rooba is completed her Ph.D in Bharathiar University, Coimbatore. She is working as a Assistant Professor in the Department of Computer Technology and Information Technology in Kongu Arts and Science College (Autonomous), Erode. She has 16 years of teaching experience. She has published 15 research papers in reputed journals.Her research interest includes Data Mining, Web Mining and Semantic Web mining.