IJCTA, 9(40), 2016, pp. 77-85
© International Science Press

An Enhanced Web Page Recommendation Technique using QTFP Growth Algorithm and Kth Markov Model

R. Suguna*

Abstract : In current world, people are busy with engagements and the customary calendar day is not sufficient for comprehensive their needs. In order to facilitate the customers, companies are rushing to post their details worldwide. So, information is overloaded in the Internet and getting the pertinent information is a complex task. In the web search framework, it is difficult to identify the user's anticipation and satisfy their requirements. Web recommendation system helps to satisfy the above addressable issues in the web search. In this paper an efficient web page recommendation technique is proposed for web page recommendation to the website visitors. To achieve an efficient web page recommendation, web logs are preprocessed using UILP algorithms and grouped using enhanced cBoid algorithm according to the association rule mining technique to find the frequently visited web pages by the users. Quality and Time based Frequent Pattern (QTFP) growth algorithm is newly introduced for effective frequent pattern identification. Kth Markov model is used to predict and recommends the web pages to the website visitors. The web logs for 98320 users are randomly generated. The algorithm is implemented in Java language. The proposed recommendation technique is compared with existing techniques with respect to precision, applicability and hit ratio.

Keywords : Web Recommendation, Association Rule Mining, Web Pages, Preprocessing, Clustering.

1. INTRODUCTION

Web recommendation is related with web personalization to carry out the needs of the customers (Mobasher et al 2000). A personalized website identifies its users and collects the information based on their expectation. It attempts to streamline the website contents depending on user's preference. Web recommendation is one of the techniques of web personalization which refers to the recommendation of a set of web pages that are coupled with the interests and preferences of the user (Dimitrios et al 2003).

Web usage mining plays a vital role to recommend the web pages to the users. User's browsing history is the base for learning the behavior of the users. Yoon (2002) and Dimitrios (2003) have mentioned that web usage mining is an important area for developing efficient recommendation systems. Web usage mining is combined with pattern mining techniques such as clustering and association rules to develop an efficient recommendation technique. Web usage mining is a proper medium to extract the users browsing history.

In this paper an efficient web page recommendation technique is proposed for web page recommendation to the website visitors. In order to achieve an efficient web page recommendation, web logs are preprocessed and grouped according to the association rule mining technique to find the frequently visited web pages by the users. Quality and Time based Frequent Pattern (QTFP) growth algorithm is newly introduced for effective frequent pattern identification. Markov model is used to predict and recommends the web pages to the website visitors. This paper deals with the proposed recommendation technique, frequent pattern

* Department of Computer Science Theivanai Ammal College for Women(Autonomous) Villupuram Tamil Nadu, India *sugunarajasekar@gmail.com*

mining algorithm, QTFP growth algorithm and Markov model for web recommendation. The performance of the proposed technique is compared with the existing techniques namely closed sequential base and weighted association rule techniques to prove its efficiency.

2. LITERATURE REVIEW

Web page recommendation system based on closed sequential pattern is developed by Utpala et al (2010). During the first step, sequential web access patterns are mined. Later more accurate patterns namely closed sequential access patterns are identified. Recommendation is provided based on pattern tree constructed with respect to closed web access patterns. A personalized web page recommendation model is proposed by the authors Qingyan et al (2010) which is based on collaborative filtering approach and Markov model. Graph based iteration algorithm is derived to identify the similar topics and recommendation is based on the probability model named Markov model.

Rana et al (2009) introduced the distributed learning automata to realize the behavior of previous users and recommend the web pages based on the learned automata. Every day huge numbers of web pages are newly added in the websites. So, this method focuses on the problem of recommending unvisited or newly added pages. Weighted association rule mining algorithm is used to address the above problem. Hits algorithm is used to extend the recommendation set. This method provided an opportunity for newly added web pages and rarely visited web pages.

The research for web page recommendation is extended by Rana et al (2009) by using weighted association rule mining concept. In this research, the association rule mining concept is extended by adding the weights for each web page visited by the user for an efficient web page recommendation. The weight is assigned to the web pages with respect to the time period spent for each web page and frequency of visit of that web page by different users. The proposed weighted association rule based recommendation technique gives improved performance than the traditional association rule mining based recommendation technique.

3. PROPOSED RECOMMENDATION TECHNIQUE

In this paper a web page recommendation technique is developed in association with quality and time based frequent growth algorithm and web usage mining. The proposed Quality and Time (QT) based recommendation technique uses pattern discovery algorithms and web usage mining. Collaborative filtering approach and pattern mining algorithms are used to make the web page recommendation more efficient. The steps involved in the proposed recommendation technique are (*i*) Data preparation using UILP algorithm (Suguna R & Sharmila D 2013a) (*ii*) Clustering the web pages for each user using enhanced cBoids algorithm (Suguna R & Sharmila D 2013b) (*iii*) Mining constraint association patterns using Quality and Time based Frequent Pattern (QTFP) growth algorithm (*iv*) Recommendation based on collaborative filtering approach and (*iv*) Recommendation of web pages using Kth-Markov model.

The frequent pattern tree is formed based on the total support values of the web pages which are calculated using the quality and time duration of the web page. The recommendation is generated by using Markov model based on the frequent pattern tree.

3.1. UILP Preprocessing Algorithm

The preprocessing weblogs are considered as the initial process which has the result of following five attributes:

<ip, user, url, session, frequency>

Where *ip* is the ip address, user is the user name, url is website address, session is time duration spent on each website by the user and frequency is the number of visits by the user (Suguna R & Sharmila D 2013a).

3.2 Enhanced cBoid Algorithm

Websites and users are grouped with respect to the parameters session time and frequency value by using enhanced cBoid algorithm (Suguna R & Sharmila D 2013b) .

4. FREQUENT PATTERN MINING ALGORITHM

Frequent pattern mining is an important data mining task and getting more attention in data mining research. Frequent pattern mining was first proposed by Agarwal & Srikanth (1994) for market basket analysis in the name of association rule mining.

In the proposed recommendation technique association rule mining concept is applied to the web log files for finding the frequently visited web pages by the website visitors. The weight of each web page is calculated by using session duration and quality rate of the web page.

5. PROPOSED QTFP GROWTH ALGORITHM

In this paper QTFP growth algorithm is used to mine the frequently visited web pages by each user using high likeness cluster. Here, the total support value for each web page is calculated to mine the frequent patterns. The total support value is calculated based on the frequency of the web page, quality rate of the web pages and time spent on the web pages. The quality rate of the web page *q* is calculated with respect to the time duration spent on each web page using the equation

$$q = (t/60)*2$$

where *q* is quality rate for the web page and *t* is time duration spent on each web page.

The total support value is calculated using the equation

$$\text{TSV} = \frac{(\text{S}_f + \text{S}_t + \text{S}_q)}{3}$$

where TSV is total support value, S_f is support value of frequency for the web page P, S_t is support value of duration of time spent for the web page *p* and S_q is support value of quality for the page *p*.

The support value of the frequency is based on the number of times that a particular user visited a particular web page. The support value of the duration of time visited for a particular web page on a particular user is calculated by dividing the duration of time spend for a web page by the maximum duration of time spend for that web page among the whole user. The support value of quality is calculated by dividing the quality value of a web page with its maximum quality value among the whole user.

The minimum support value (10, 20, 30, 40 and 50) is set and compared it with TSV of the web pages. From the result of calculation the web pages which have the total support value more than the minimum support value are taken into consideration and arranged in descending order for each user.

5.1. Frequency Database Generation

The database is generated based on the web pages which the users visited and the time duration which the users spend to view those web pages and the quality of those web pages.

This database contains two sections, the first section contains the web pages which are visited by the users and the second section contains the time duration and quality of the web pages based on the first section.

$$\text{BFD}_i = \{p_1, p_2, ..., p_n\}$$
$$\text{BFD}_j = \{t_1, q_1, t_2 q_2, ..., t_n q_\text{n}\}$$

Where *p* is web pages, *t* refers duration of time spent for the web page *p* and *q* is quality rate for each web page given by the users.

The following example shows the sample database which contains the users and the web pages visited by them.

Example 1 : Web pages visited by the users

Table 1

Users	*Web Pages Visited*
U1	*w, n, b, j, o, h*
U2	*n, s, b, w, k*
U3	*s, w, m, c*
U4	*s, b, i, g*
U5	*n, w, b, d, k, v*

In the above example, the user U1 has the web pages *w, n, b, j, o* and *h*. These web pages are extracted from the high likeness clusters using proposed cBoids and UCC algorithms. The extracted web pages are considered as most frequent and most likely web pages by the user U1. In the same way the user U2 likes the web pages *n, s, b, w* and *k*. The web pages *s, w, m* and *c* are frequently visited by the users U3. The user U4 visits the web pages *s, b, i* and *g* respectively. The user U5 visits the pages *n, w, b, d, k* and *v* respectively.

The parameters considered to generate the frequency database is duration of time spend on a particular web page by the user and the quality of the web page. The duration of time spend on a particular web page by a user is one of the important field to identify the preference of the web page. Because if a web page has more importance then it is assumed that the user accesses it for a long time and it shows the interest of that user on that particular web page.

Another parameter which is considered to generate the frequency database is quality rate of the web page. This is also an important factor which is considered to generate the biased frequency database. So, these two parameters are applied to generate the frequency database effectively.

Table 2 shows the time duration and quality rate of the web pages which are mentioned in example 2 are considered. The user U1 spent 60 minutes for web page w and the quality rate for the web page is 2. The time duration spent for the next web page n is 30 minutes and the quality rate for the web page is 1. The third web page visited for the user U1 is b and the time duration spent for that web page is 140 minutes and quality rate for the web page is 4.6. In the same manner the database are generated for web pages which are visited by the user, time duration spent for the web pages and quality rate of the web pages.

5.2. Frequent Pattern Tree Generation

Example 2 shows the sample web pages which have the total support value as high compared to the minimum support value and arranged in descending order for each user based on the total upport value.

Table 2
Sample Database with Time Duration and Quality

Users	*(Time Duration, Quality)*
U1	(60,2), (30,1),(140,4.6),(180,6),(130,4.3),(20,0)
U2	(50,1.6), (80,2.6), (100,3.3), (110,3.6),(120,4)
U3	(10,0),(40,1.3),(60,2),(180,6)
U4	(80,2.6),(130,4.2),(120,4),(180,6)
U5	(140,4.6),(130,4.3),(100,3.3),(10,0),(20,0),(110,3.6)

Example 2 : Web pages which have high total support value than the minimum support value

Table 3

Users	*Web Pages Visited*
U1	b, o
U2	s, b, w
U3	s, w, m
U4	s, b, i
U5	b, w, d

5.3 Conditional Pattern Base Generation

The conditional pattern base is generated for mining the FP tree to extract the needed information from the tree. Table 4 depicts the conditional pattern base for each web page.

Table 4

Nodes with Conditional Frequent Patterns and Values

Node	*Conditional Frequent Patterns with Values*
m	(m = 1), (sm = 1), (wm = 1)
i	(i = 1), (si = 1), (bi = 1)
w	(w = 3), (sw = 2), (bw = 2)
d	(d = 1), (bd = 1), (wd = 1)
o	(o = 1), (bo = 1)
b	(b = 4), (sb = 2)
s	(s = 3)

The web page *b* has highest weight among all the web pages because it is visited 4 times by the users. The web page sequence (*s*, *b*) is visited 2 times. The web page *s* and *w* has next highest visit, it has the weight 3.

6. WEB PAGE RECOMMENDATION USING MARKOV MODEL

The basic procedure of Markov model is to predict the next action based on the previous action. Markov model is extensively used to recommend the web pages based on the past browsing history of the users. It generates recommendation of web pages only by assumption with respect to previously visited web pages by the users (Magdalini et al 2005). Higher order models improve the efficiency of web page prediction but it leads time complexity. In web page prediction problems the Kth-Markov model has the probability function of determining Kth visited web page is based on the k-1 ordered sequence of web pages visited by the users.

In the proposed web page recommendation technique the Kth-Markov model (where $k = 3$) (Mamoun & Issa 2012) is used. When the web pages are recommended for new user, the sequence path of that user is compared with the FP tree and it recommends the web pages using the probability definition.

Consider the input sequence of web pages which the user given as $p_1, p_2, \ldots, p_n$ where $p_1, p_2, \ldots, p_n$ are the sequence of web pages visited by the new user. Initially the input sequence obtained from the user is compared with the FP tree and provides the three best matching results represents in equations

$$p_1, p_2, \ldots p_n, p_{n+1} \ldots, p_k \ (5.6)$$
$$p_1, p_2, \ldots p_n, p_{n+1} \ldots, p_k \ (5.7)$$
$$p_1, p_2, \ldots p_n, p_{n+1} \ldots, p_l \ (5.8)$$
$$p_1, p_2, \ldots p_n, p_{n+1} \ldots, p_m \ (5.9)$$

Secondly the total support value for each sequence is calculated. The important point is, every sequence has separate total support value and the most important sequence is recommended to the user by calculating the probability.

$$p_{n+1}^{(1)} = \text{Pro } o(p_1, p_2, \ldots\ldots, p_n, p_{n+1}, \ldots\ldots, p_k / p_1, p_2, \ldots\ldots, p_n)$$
$$p_{n+1}^{(2)} = \text{Pro } o(p_1, p_2, \ldots\ldots, p_n, p_{n+1}, \ldots\ldots, p_l / p_1, p_2, \ldots\ldots, p_n)$$
$$p_{n+1}^{(3)} = \text{Pro } o(p_1, p_2, \ldots\ldots, p_n, p_{n+1}, \ldots\ldots, p_m / p_1, p_2, \ldots\ldots, p_n)$$

The above probability Pro(P_{n+1}/P) is calculated by using the equations based on the sequence of users in the FP tree constructed for 50000 users from frequency database in 0.4 seconds.

$$\text{P}(p_{(n+1)}^{(1)} / \text{ P} = \frac{\text{T}_{\text{sup}}(p_1, p_2, \ldots, p_n)}{\text{T}_{\text{sup}}(p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_k)}$$

$$\text{P}(p_{(n+1)}^{(2)} / \text{ P} = \frac{\text{T}_{\text{sup}}(p_1, p_2, \ldots, p_n)}{\text{T}_{\text{sup}}(p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_l)}$$

$$\text{P}(p_{(n+1)}^{(3)} / \text{ P} = \frac{\text{T}_{\text{sup}}(p_1, p_2, \ldots, p_n)}{\text{T}_{\text{sup}}(p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_m)}$$

The final recommendation is based on the equation

$$p_{n+1} = \text{arg max } (p_{n+1}^{(1)}, p_{n+1}^{(2)}, p_{n+1}^{(3)})$$

7. PERFORMANCE ANALYSIS

Data preprocessing and clustering processes are carried out in the synthetic data sets using UILP preprocessing algorithms and proposed cBoids algorithm. High likeness cluster for each user is obtained for generating QTFP tree. Markov model is applied in association with QTFP tree to generate recommendations. Web logs for 98320 users are randomly generated and considered for evaluation. The input data sets are divided into two groups (*i*) Training set which are used to generate the frequent pattern tree with respect to quality rate, time duration and frequency of the web pages (*ii*) Testing data set which is used for evaluating the performance of web page recommendation. Frequent tree is constructed for 50000 users and remaining 48320 users are considered for evaluating proposed recommendation technique. The system is limited to recommend n web pages where $n < 10$.

7.1. Performance Metrics

In this paper, the performance evaluation is carried out by QT based recommendation system associated with Markov model is evaluated against existing recommendation.

The performance of the proposed recommendation technique is evaluated in terms of precision, applicability and hit ratio.

Precision : Precision defines the measure of correct recommendation among the correct and incorrect recommendations generated by the system.

$$\text{Precision} = \frac{\text{C}^+}{\text{C}^+ + \text{I}^-}$$

Applicability : Applicability defines the measure of prediction of correct and incorrect recommendation among total number of request by the system. The term applicability is defined as the sum of correct recommendation and incorrect recommendations divided by total number of given request.

$$\text{Applicability} \quad = \quad \frac{C^{+} + I^{-}}{|\,N\,|}$$

Hit ratio : Hit ratio predicts the overall performance of the system. It is the product of precision and applicability.

$$\text{Hit ratio} = \text{Precision} \times \text{Applicability} \quad = \quad \frac{C^{+}}{|N|}$$

where C^{+} denotes the number of correct recommendations, I^{-} denotes the number of incorrect recommendations and |N| denotes the total number of given requests.

7.2. Performance Comparison of QT and Markov Model Based Recommendation Technique with Existing Techniques

The performance of the proposed QT and Markov model based recommendation technique is compared against Closed Sequential Pattern technique and Weighted Association Rule technique. The performance of both the techniques is compared in terms of precision, applicability and hit ratio. These parameters are evaluated with different minimum support values.

Table 5 consolidates the performance of algorithms with the parameters precision, applicability and hit ratio. From the analysis carried out in three different algorithms it is observed that hit ratio for all the different level of minimum supports. The minimum support values 10, 20, 30, 40 and 50 are assigned for performance evaluation. the QT and Markov model based recommendation technique gives higher precision, applicability and hit ratio.

Table 5

Comparison of Web Page Recommendation Techniques

Recommendation Techniques	*Precision in %*					*Applicability in %*					*Hit Ratio in %*				
	Minimum Support					*Minimum Support*					*Minimum Support*				
	10	*20*	*30*	*40*	*50*	*10*	*20*	*30*	*40*	*50*	*10*	*20*	*30*	*40*	*50*
Closed Sequential Pattern Technique	75.15	72.13	68.56	59.42	52.69	90.36	88.26	85.47	76.83	71.58	67.91	63.66	58.60	45.65	37.72
Weighted Association Rule Technique	85.65	82.45	82.45	75.82	74.83	94.24	92.25	92.25	90.84	90.04	80.72	76.06	76.06	68.87	67.38
QT and Markov Model Base Technique	88.75	88.86	88.35	85.43	85.43	96.58	96.62	95.89	93.65	93.65	85.71	85.86	84.72	80	80

Figure 1 depicts the performance comparison of recommendation techniques in terms of precision. From the figure it is observed that, Weighted Association Rule recommendation technique gives precision value as 85.65% for minimum support 10, 82.45% for minimum support 20 and 30. If the minimum support values are set to 40 and 50, it produces the precision values 75.82% and 74.83% respectively. The Closed Sequential Pattern technique gives better precision for the minimum support 10 and 20.

The proposed QT and Markov model based recommendation technique gives the precision value for the minimum support 10 as 88.75%. When the minimum support level is increased as 20, it gives 88.86%. For the minimum supports 30, 40 and 50, the proposed technique gives 88.35%, 85.43% and 85.43% respectively.

The performance in terms of applicability is depicts in Figure 2. The applicability values for QT and Markov model base technique are 96.58%, 96.62%, 95.89%, 93.65% and 93.65% respectively for different minimum support value. Weighted Association Rule technique gives the applicability values for various minimum supports are 94.24%, 92.25%, 92.25%, 90.84% and 90.04% respectively. When the minimum support level is increased, the applicability is decreased. The applicability of Closed Sequential

Base technique for specified minimum support values are 90.36%, 88.26%, 85.47%, 76.83% and 71.58% respectively. There is a decrease in applicability value when the minimum support is increased.

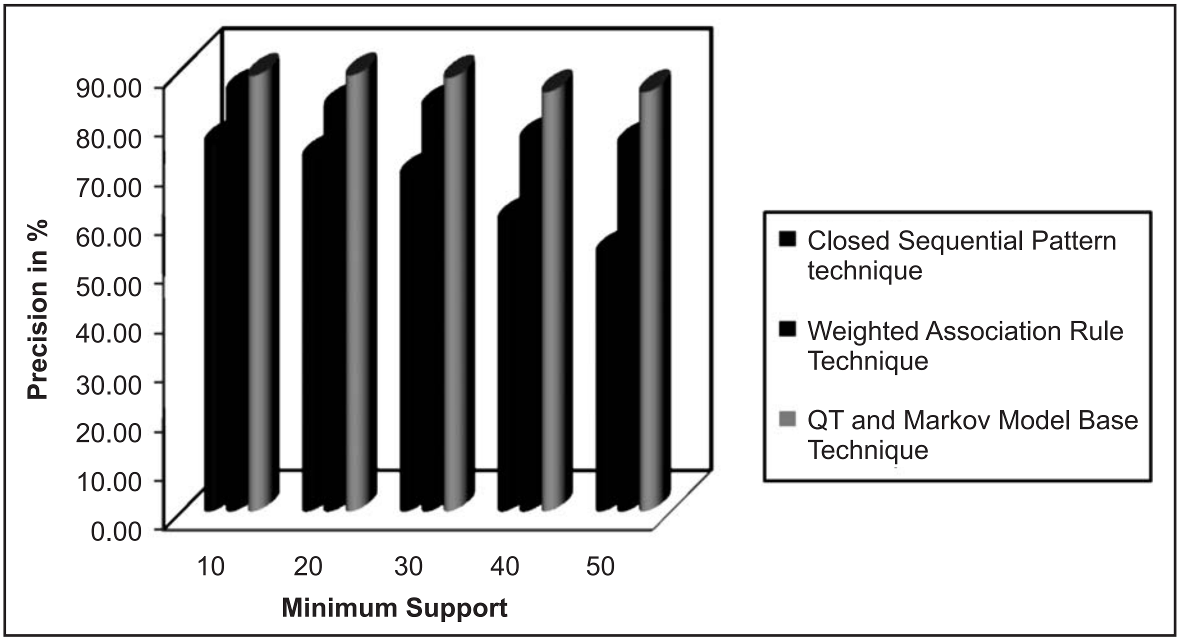


Figure 1: Performance Comparison in terms of Precision

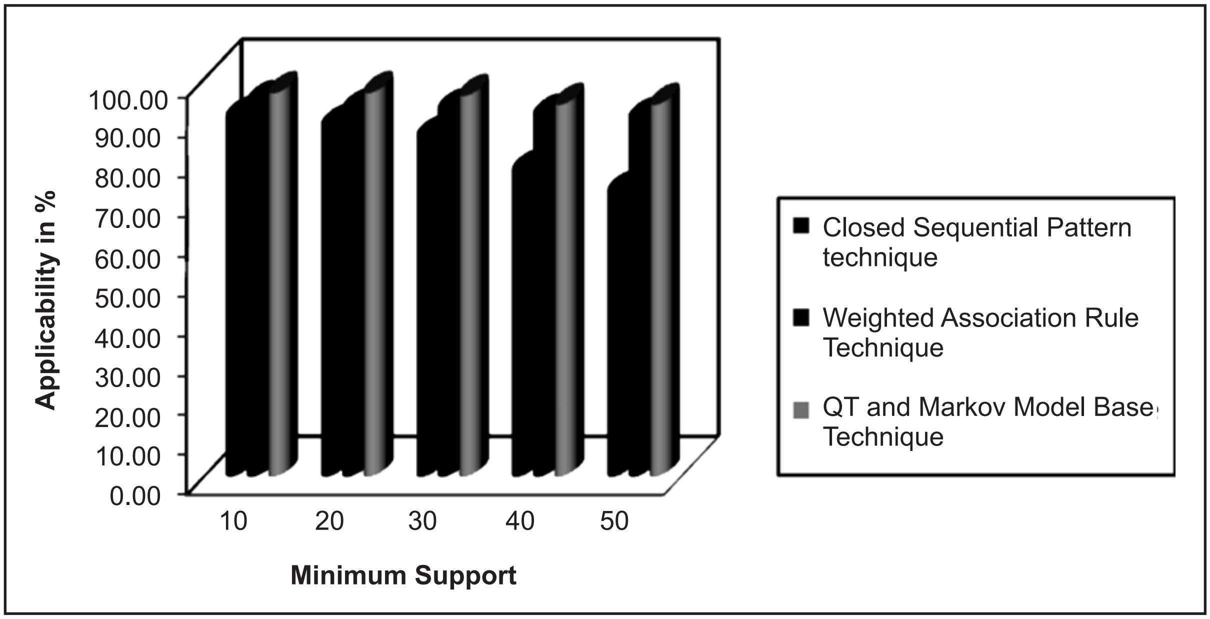
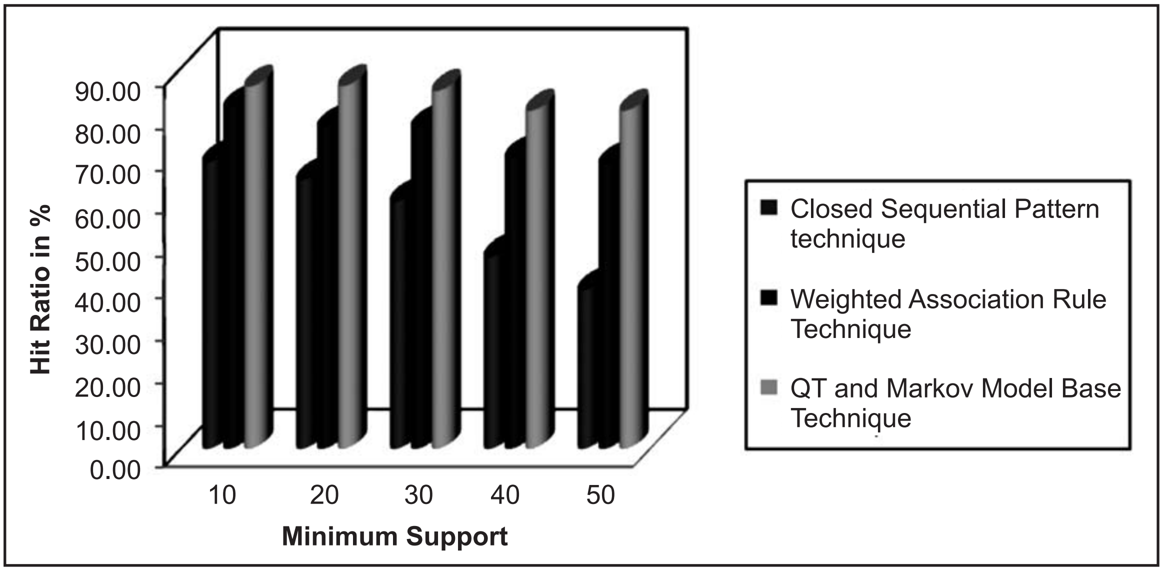


Figure 2: Performance Comparison in terms of Applicability

Figure 3 shows the performance comparison between QT and Markov model base technique, Closed Sequential Base technique and Weighted Association Rule technique with respect to hit ratio. Different minimum support values are used to evaluate the performance in terms of hit ratio. When minimum support value is set as 10, the hit ratio of QT and Markov model base technique is 85.71%, the Weighted Association Rule technique gives 80.72 and the hit ratio of Closed Sequential Base technique is 67.91%. For the minimum support value 20, the hit ratio is 85.86% for QT and Markov model base technique, 76.06% for Weighted Association Rule technique and it is 63.66% for Closed Sequential Base technique. When the minimum support value is 30, the hit ratio is 84.72% for QT and Markov model base technique, 76.06% for Weighted Association Rule technique and the hit ratio is 58.60% for Closed Sequential Base technique. The hit ratio obtained for Weighted Association Rule technique is 68.87% and 67.38% for the minimum support values 40 and 50.



Figur 3: Performance Comparison in terms of Hit Ratio

8. CONCLUSION

Web logs are preprocessed and clustered according to QTFP growth algorithm. Recommendation technique is proposed based on the frequent patterns. The recommendation accuracy and its performance are compared with the existing algorithms namely Closed Sequential Pattern technique and Weighted Association Rule technique with the measures precision, applicability and hit ratio. The experimental result shows the improved performance of the proposed system for web page recommendation than the existing techniques.

9. REFERENCES

1. Agrawal, R &Srikant, R 1994, 'Fast Algorithms for Mining Association Rules in Large Databases', Proceedings of the 20th International Conference on Very Large Data Bases, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, pp. 487-499.
2. Dimitrios, P, Georgios, P, Christos, P & Constantine, DS 2003, 'Web Usage Mining as a Tool for Personalization: A Survey', User Modeling and User Adapted Interaction, vol. 13, no. 2, pp. 311-372.
3. Magdalini, E, Michalis, V & Dimitris, K 2005, 'Web Path Recommendations based on Page Ranking and Markov Models', Proceedings of the 7th annual ACM international workshop on Web information and data management,WIDM'05, pp. 2-9.
4. Mobashar, B, Cooley, R & Srivastava, J 2000, 'Automatic Personalization Based on Web Usage Mining', Communications of the ACM, vol. 43, no. 8, pp. 142-151.
5. Mamoun, A & Issa, K 2012, 'Prediction of User's Web-Browsing Behavior: Application of Markov Model', IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, vol. 42, no. 4, pp. 1131-1142
6. Mobashar, B, Cooley, R & Srivastava, J 2000, 'Automatic Personalization Based on Web Usage Mining', Communications of the ACM, vol. 43, no. 8, pp. 142-151.
7. Rana, F, Mohammad, RM & Afsaneh, R 2009, 'An Efficient Algorithm for Web Recommendation Systems', IEEE International Conference on Computer Systems and Applications, pp. 579-586.
8. Rana, F, Meybodi, MR & Ghari, AN 2009, 'Web Page Personalization based on Weighted Association Rules', IEEE International Conference on Electronic Computer Technology, pp. 130-135.
9. Utpala, N, Subramanyam, RB & Khanaa, V 2010, "Developing a Web Recommendation System Based on Closed Sequential Patterns", Communications in Computer and Information Science, vol. 101, no. 1, pp. 171-179.
10. Suguna, R & Sharmila, D 2013a, 'User Interest Level based Preprocessing Algorithms using Web Usage Mining', International Journal of Computer Science and Engineering (IJCSE), vol. 5, no. 9, pp. 815-822. (ISSN : 0975-3397).
11. Suguna, R & Sharmila, D 2013b, 'Enhanced cBoids Algorithm for Web Logs Clustering', International Journal of Engineering Trends and Technology (IJETT), vol. 6, no, 4, pp. 189-197. (ISSN : 2231-5381).
12. Yoon, HC, Jae, KK& Soung, HK 2002, 'A Personalized Recommender System based on Web Usage Mining and Decision Tree Induction', Expert Systems with Applications, vol. 23, pp. 329-342.