Personalized Recommendation of Web Pages Using Group Average Agglomerative Hierarchical Clustering (GAAHC)

Harish Kumar B T
Dept of CSE
Bangalore Institute of Technology
Bangalore
India
harish.bitcse82@gmail.com

Dr. Vibha L
Dept of CSE
B.N.M Institute of Technology
Bangalore
India

Dr. Venugopal K R,
University Visvesvaraya College of Engineering
Bangalore
India

Abstract:-Entrepreneurs are investing heavily on marketing and promoting business through the websites to enhance their online reputation and draw the attention of the web users. Website structure plays the vital role in attracting the web users. Creating personalized website structure for individual user by restructuring the web site structure is a tedious and endless job. If the users do not find the required information easily in the websites, then users abandon such websites. Hence, personalized recommendation of web pages to the web users increases the user's interest and the time they spend in the website. Personalization is the process of creating customized participation of users to a website, rather than providing a broad participation. Personalization allows the website to present the users with the unique participation bespoke to their demands and passion. Personalized recommendation is a challenging task, which has drawn the focus of many researchers. Personalization has to trace the behavior of individual users. Usage behavior can be traced by observing the individual navigation patterns using web log file of the specific website. This method requires session identification, clustering sessions into similar clusters and building a model for personalized recommendations using access time length and frequency of access. Most of the existing works on this topic have used K-Means clustering with Euclidean distance. K-Means suffers from choosing the initial random center and sequence of page visits is not considered. The proposed research work uses Group Average Agglomerative Hierarchical Clustering (GAAHC), with Modified Levenshtein Distance (MLD) and page rank using access time length and the frequency of page access.

Key Words: Personalization, Recommendation, Agglomerative, Clustering, Levenshtein.

1. Introduction

The plenty of data available on World Wide Web has captured the users to explore information via internet. This plethora of information often creates problems with users unable to obtain useful and relevant information. The website user's demands have shifted towards the personalized participation in the websites. Personalized recommendation of web pages typically includes the challenges like i) An over-abundance of non actionable data ii) Knowing who to personalize for iii) Measuring the impact of personalization. Website personalization can be accomplished with the following strategies.

1. **Finding Audiences:** This can be done by targeting visitors to websites and based on the actions performed by them
2. **Planning personalized recommendation:** understanding visitors will help in planning the recommendation to website tailored to specific audiences.
3. **Continuous measurement and improvement:** Constantly measuring the return on website personalization is necessary as not every personalized recommendation will resonate (sound). So, it is necessary to understand how these personalized recommendations are used and adjust accordingly.

The above strategies are performed by considering the web log data of NASA website. The web log data is in Common Log Format (CLF). Legacy web server software records the user access to website using CLF format. CLF carries the following fields.

1. Client machine IP address
2. User Identifcation
3. Web Page access date and access time
4. HTTP request method (GET/POST)
5. Relative path of the web page on the server
6. Transmission Protocol
7. HTTP status code
8. Total number of bytes transmitted

Eg: 205.212.116.107 - - [02/Aug/1995:00:02:12 -0400] "GET /shuttle/CountDown/CountDown.html HTTP/1.0" 200 4985.

1.1 Motivation

Website users have the tendency to leave the website if they fail to find the required information from the website. Hence personalization is a powerful way to communicate speculatively with web users and recommend the web pages to their particular needs. Personalization strategy allows you to identify clusters of website users with distinct preferences and then create personalized recommendations to them. Personalized recommendation of web pages helps the web users to find the required and useful information in less time. Reduces probability of web user getting distracted from using the website and also enhances the website user's interest in using the website for longer period of time.

1.2 Contribution

In the present work, personalized recommendation of web page is done based on usage pattern using GAAHC with modified Levenshtein distance and web page ranking with access duration and access frequency. In the first part, web log data is pre-processed; user and session clusters are formed using GAAHC with modified Levenshtein distance. In the second part, page rank is computed for the web pages within each cluster at every level of hierarchy using access time length and access frequency. In the third part, personalized recommendations are generated to individual users.

The remaining sections of the paper are arranged as given below. Overview of the related work in the proposed research work is presented in section II. Problem statement, aim and objectives are discussed in section III. Architecture of the proposed work is presented in section IV. Section V covers proposed methodology and working examples. Section VI discusses experimental setup and performance analysis. Finally, section VII concludes the proposed research work.

2 Related Work

This section provides a short literature review on the latest research works done related to this research area. Literature review is useful in studying and comparing the available preprocessing and clustering techniques.

Hiral Y. Modi in [1] has presented a Hybrid Clustering Using K-Means algorithm and Pattern Matching approach using Boyer Moore algorithm for recommending the products to the online users based on users' recent transaction history. K-Means algorithm has the drawback of choosing the initial random **centroid**. K-means clustering does not take the page visit sequence into consideration.

In [2] A Vinupriya et al., have proposed web page personalization and prediction of link using inverted index and FLAME clustering. This work is based on many parameters like hit count of web pages, access length of every web pages, download count and link distance. This approach has used the generalized inverted index frame work for quick result discovery. An efficient ranking of web pages based on web page relevance and personalized search was proposed by Mercy Paul Selvan in [3]. This work is based on the user's feedback and the Markov model to find the importance of the page and user profile for personalization.

Zhongyun Ying in [4] has proposed an algorithm for personalized web page recommendation using the improved collaborative algorithm to find web page recommendation sets with similar users with the interested pages. Web pages recommendation sets are merged using Merge sort algorithm (MSA). This work does not focus on the continuous measurement to find whether the recommendation sounds to the users or not. Gerrad Deepak et al., in [5] have proposed a differential semantic algorithm for query relevant web page recommendation. This work computes the semantic similarity between the query words, keywords (title tag words and URL words) and content words (body tag words) using the Adaptive Pointwise Mutual Information (APMI) strategy. The proposed methodology extracts the query relevant URLs using the query words. Keywords are extracted from the URLs, title tag and content words are extracted from the body tag of HTML.

Korinna Bade in [6] has proposed a personalized hierarchical clustering of web pages specific to user web search results. Dipa Dixit in [7] has proposed a two tier architecture to capture page visits made by users in the form of recommendation list and also lists the pages visited by other users having similar usage profile. In [8] K. Suneetha has discussed about the techniques for analyzing the performance of web page recommendation using Markov model and weighted sequential patterns. This work proposes changes to the traditional sequential pattern mining by incorporating measures like time spent and latest view to mine more useful patterns.

Neeraj Iyer in [9] has conducted a survey on online recommendation using web usage mining and tried to make a comparative study of the techniques which were used in the previous work. V Chitraa in [10] has proposed a new technique for recommending online users the web pages using web log data. This technique uses fuzzy clustering with Euclidean distance and Longest Common Subsequence (LCS) Algorithm to classify users.

Zohreh Anari in [11] has presented how to determine the similarity between the web pages using Learning Automata (LA) and Hypertext Probabilistic Grammar (HPG). LA is a decision making technique which works in unknown random environments for learning the optimal action from the set of available actions. HPG has one-to-one mapping between a non-terminal symbols set and terminal symbols set. Non-terminal symbols correspond to web pages and terminal symbols called as production rule correspond to the link between the pages. S. Ramanamurthy in [12] has suggested the idea of page ranking using genetic algorithm. The algorithm finds the synonyms for the keywords using wordnet, opens the URL and check whether the keyword and the similar words appear in that webpage to compute the page rank. In [13] Harish Kumar B T et al., has proposed a Single Link Hierarchical Clustering using Modified Levenshtein Distance to predict the web page access in the future by the user.

3 Problem Statement

Website Users leave the websites if they fail to obtain the required useful information. Given the Web Log document D, Let $U=\{U_1,U_2,\ldots U_n\}$ be the distinct users, $P=\{P_1,P_2,\ldots P_n\}$ be the distinct web pages, Let $S=\{S_1,S_2,\ldots S_n\}$ be the web sessions for all users U_i, where each session S_i is an ordered sequence of pages P_i accessed by the user U_i. Let $C=\{C_1,C_2,\ldots C_n\}$ be the set of clusters with L levels formed using GAAHC. Page Rank $PR=\{PR_1,PR_2,..PR_n\}$ is computed for each page $P_i \in C_i$ at each cluster level L_i using access time T_i and access frequency F_i for the Web page P_i. Then recommendation $R=\{R_1,R_2,\ldots R_n\}$ for the users U_i is generated.
Objectives:

1. Preprocessing the web log.
2. Computing the similarity of user sessions and cluster forming using GAAHC.
3. Ranking the web pages in cluster at level $L_{i.}$
4. Recommending based on page rank.

4 Architecture and Modeling

The proposed work uses the NASA web log file as input. The general architecture and modeling of the proposed system is illustrated in Fig.1. The NASA web log file is preprocessed to remove entries with *.gif, *.jpg, *.css, 404 and 500 status code entries. Distinct users and distinct web pages are indentified. User and session identification is done then the user and session are clustered with GAAHC using modified Levenshtein distance. Page rank for every page in the cluster at level L_i is computed using the access time length of page and frequency of access of the pages. Personalized recommendation system generates the personalized web pages to the users using the page rank.

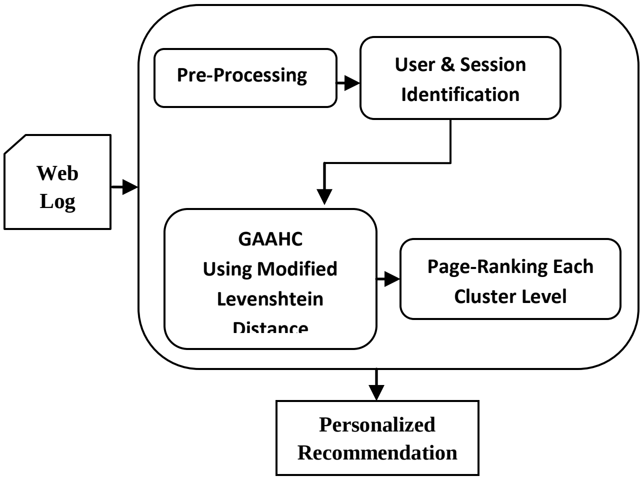
Web Log
Pre-Processing
User & Session Identification
GAAHC Using Modified Levenshtein Distance
Page-Ranking Each Cluster Level
Personalized Recommendation

Fig. 1: Architecture and Modeling

5 Proposed Methodology

The proposed methodology intent is to improve the accuracy of the personalized recommendation. The proposed methodology uses the MLD shown in equation (1) to find the similarity between the user sessions. Cluster formation is done based on the GAAHC using the equation shown in (2).

$M(i, 0) = 0, 0 \leq i \geq L1$

$M(0, j) = 0, 0 \leq j \geq L2$

Where,

$M(i, j)$ is maximum similarity score matrix.

$L1$ and $L2$ are lengths of sessions S_i and S_j respectively.

$$M(i, j) = Max\begin{cases} 0, \\ \mathrm{M(i-1, j-1) + 2\ if\ S_i = S_j\ else -1,} \\ \mathrm{M(i-1, j) + W_{d,}} \\ \mathrm{M(i, j-1) + W_{i,}} \end{cases}$$

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

For $1 \leq i \geq L1, 1 \leq j \geq L2$

Where,

W_i = -1 (insertion penalty) and W_d = -1 (deletion penalty)

The Group Average Similarity between any two clusters C_i and C_j is the average similarity between any object $O_i \in C_i$ and any object $O_j \in C_j$.

$$GA-SIM\left(C_i, C_j\right) = \frac{1}{\left|C_i\right| \times \left|C_j\right|} \sum_{O_i \in C_i, O_j \in C_j} d(O_i, O_j)$$

----------------- (2)

The maximum similarity between all the user session pair is computed by using the MLD shown in equation (1). Table I shows the example session patterns used for the computation. Maximum similarity computation between session pair (S1, S2) is shown in example 1. The maximum entry in the example1 is at *M(4, 6) = 6*. The max similarity between session *S1* and session *S2* is 0.75, computed using the equation (3). Table II shows the max similarity between every pair of sessions depicted in Table I.

TABLE I: Session Access Pattern

User-ID	Session-ID	Access Pattern
U1	S1	A, B, C, D, E
U2	S2	C, D, E
	S3	B, C, D
U1	S4	B, C
	S5	A, B
U3	S6	A, B, D, E

Example 1: Max Similarity computation between Session S1 and S2

	S1 Access Pattern						
		---	A	B	C	D	E
S2 Access Pattern	---	0	0	0	0	0	0
	C	0	0	0	2	1	0
	D	0	0	0	1	4	3
	E	0	0	0	0	3	6

$$MAX-SIM=\frac{Max\{M(i,j)\}}{Li+Lj}$$

--------------- (3)

Where,

$Max\{M(i,j)\}$ → Max element in $M(i,j)$

L_i →Length of session S_i

L_j → Length of session S_j

TABLE II: Max similarity between every pair of sessions

	S1	S2	S3	S4	S5	S6
S1	-	0.75	0.75	0.57	0.57	0.77
S2		-	0.66	0.40	0.00	0.57
S3			-	0.80	0.40	0.42
S4				-	0.50	0.33
S5					-	0.66
S6						-

GAAHC

First iteration every session is assumed as a cluster of their own session. In second iteration session *S3* and *S4* are grouped into cluster *C1= {S3, S4}* at 80% similarity level or 20% dissimilarity level as 0.80 is the maximum entry in Table II. Table III shows the group average computation between the cluster *C1= {S3, S4}* and other sessions *S1*, *S2*, *S5* and *S6* using equation (2).

Table III: Group average computation between (S1, S2, C1, S5, S6)

	S1	S2	C1	S5	S6
S1	-	**0.75**	0.66	0.57	0.77
S2		-	0.53	0.00	0.57
C1				0.58	0.37
S5				-	0.66
S6					-

In third iteration session *S1* and *S2* are grouped into cluster *C2 = {S1, S2}* at 75% similarity level or 25% dissimilarity level as 0.75 is the maximum entry in Table III. Table IV shows the group average computation between clusters *C1*, *C2*, session *S5* and session *S6*. Session *S6* is grouped into Cluster *C2* at 67% similarity level or 33% dissimilarity level forming cluster *C3 = {S1, S2, S6}*. Table V shows the group average computation between clusters *C1*, *C3* and session *S5*. Cluster *C1* and *C3* grouped into cluster *C4 = {S3, S4, S1, S2, S6}* at 52% similarity or 48% dissimilarity level. Table VI shows the group average computation between cluster *C4* and session *S5*. Session *S5* joins the cluster *C4* at 42% similarity or 58% dissimilarity level to form cluster *C5 = {C4, S5}*. Table VII shows the summary of the hierarchical cluster formation at different levels. Dendogram representation for the clusters *C1*, *C2*, *C3*, *C4* and *C5* is shown in Fig. 2. Sessions are plotted on *X* axis and dissimilarity is plotted on *Y* axis.

Table IV: Group average computation between (C2, C1, S5, S6)

	C2	C1	S5	S6
C2	-	0.59	0.28	**0.67**
C1		-	0.45	0.37
S5			-	0.66
S6				-

Table V: Group average computation between (C3, C1, S5)

	C3	C1	S5
C3	-	0.52	0.41
C1		-	0.45
S5			-

Table VI: Group average computation between (C4, S5)

	C4	S5
C4	-	0.42
S5		-

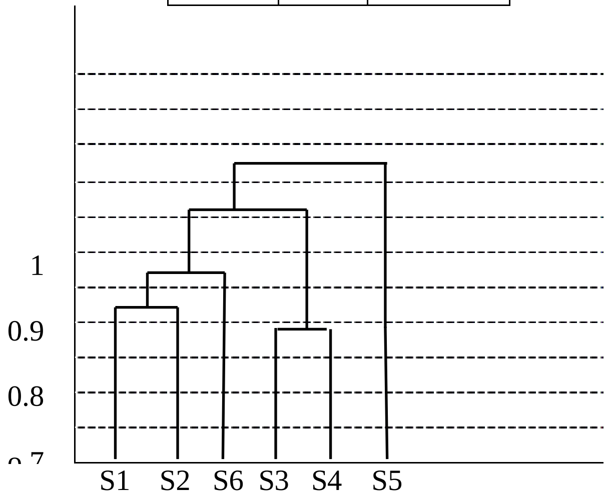

Fig. 2: Dendogram Representation

Table VII: Summary of Cluster formation

Dissimilarity	Level	Cluster-ID	Session-ID
0.20	L1	C1	{S3,S4}
0.25	L2	C2	{S1,S2}
0.33	L3	C3	{S1,S2},{S6}
0.48	L4	C4	{S1,S2},{S6}, {S3,S4}
0.58	L5	C5	{S1,S2},{S6}, {S3,S4},{S5}

Personalized recommendations are generated by constructing the Transition Probability Matrix (TPM) using the hierarchical cluster, at level L_i where user U_i all sessions $S_i \in C_i$ using equation (4). For users *U1, U2* and *U3* TPM is computed at level *L5*, *L4* and *L3* respectively is depicted in Table VIII. Page Rank PR_i of a page is computed using the equation (5). Page recommendation R_i for user U_i is done by using the equation (6).

$$P[i,j] = n[i,j] \; / \; \sum_{j=0}^{n} n[i,j] \text{ ----------- (4)}$$

n[i,j] indicates the number of transition from $page_i$ to $page_j$. P[i,j] is the probability of transition from $page_i$ to $page_j$.

Where, $P[i,j] \geq 0 \; and \; \sum_{j=0}^{\infty} P[i,j] \; = 1 \; \forall \; i,j$

$$PR_i = \sum_{i=0}^{n} TPM(i,j) \text{ ------------------ --- } (5)$$

$$R_i(U_i) = Max\{ \forall \; PRi \in \; U_i\} \text{ --------------- (6)}$$

TABLE VIII: TPM of U1, U2 and U3

User U1					
Cluster at Level: L5 **Session:** {S1, S2}, {S6}, {S3,S4}, {S5} **Page Sequence:** A B C D E C D E A B D E B C D B C A B					
	A	B	C	D	E
A	-	1.00	0.00	0.00	0.00
B	0.00	-	0.60	0.20	0.00
C	0.00	0.00	-	0.75	0.00
D	0.00	0.25	0.00	-	0.75
E	0.00	0.00	0.33	0.00	-
Rank (PR_i)	**0.00**	**1.25**	**0.93**	**0.95**	**0.75**
Recommendation $R_i(U_i)$	**B, D, C, E, A**				
User U2					
Cluster at Level: L4 **Session:** {S1,S2}, {S6}, {S3,S4} **Page Sequence:** A B C D E C D E A B D E B C D B C					
	A	B	C	D	E
A	-	1.00	0.00	0.00	0.00
B	0.00	-	0.75	0.25	0.00
C	0.00	0.00	-	0.75	0.00
D	0.00	0.25	0.00	-	0.75
E	0.00	0.33	0.33	0.00	-
Rank (PR_i)	**0.00**	**1.58**	**1.08**	**1.00**	**0.75**
Recommendation $R_i(U_i)$	**B, C, D, E, A**				
User U3					
Cluster at Level: L3 **Session:** {S1,S2}, {S6} **Page Sequence:** A B C D E C D E A B D E					
	A	B	C	D	E
A	-	1.00	0.00	0.00	0.00
B	0.00	-	0.50	0.50	0.00
C	0.00	0.00	-	1.00	0.00
D	0.00	0.00	0.00	-	1.00
E	0.33	0.00	0.33	0.00	-
Rank (PR_i)	**0.33**	**1.00**	**0.83**	**1.50**	**1.00**
Recommendation $R_i(U_i)$	**D, B or E, C, A**				

6 Experimental Setup and Performance Analysis

The proposed research work is implemented using java tries to improve the accuracy of the web page recommendation. NASA web access log from 01/July/1995 to 05/July/1995 (five days) is used as the data set for training. NASA web access log of 06/July/1995 (one day) is used to test and evaluate the proposed methodology performance. Training data set consisted of 3,79,582 records, reduced to 80,329 records after filtering and preprocessing. Reduced data set consists of 19,571 unique users and 562 unique pages. Fig. 3 depicts the number of user sessions for the different session timeout period with minimum of five page views in a session. In the proposed work session timeout is set to 30 minutes with minimum of 5 page views in a session. Table IX shows comparison between different clustering technique and distance used. Fig. 4 shows the accuracy comparison between the K-Means using Euclidean Distance (KM_ED) and Single Link Agglomerative Hierarchical Clustering using Levenshtein Distance (SLAHC_LD). Fig. 5 shows the accuracy comparison between SLAHC_LD, Single Link Agglomerative Hierarchical Clustering using Modified Levenshtein Distance (SLAHC_MLD) and Group Average Agglomerative Hierarchical Clustering using Modified Levenshtein Distance (GAAHC_MLD).

TABLE IX: Comparison of Clustering Techniques

SL. No	Clustering Technique	Distance Measure	No. of Clusters	No. of Levels
1	KM	ED	NA	NA
2	SLAHC	LD	600	115
3	SLAHC	MLD	524	93
4	GAAHC	MLD	450	80

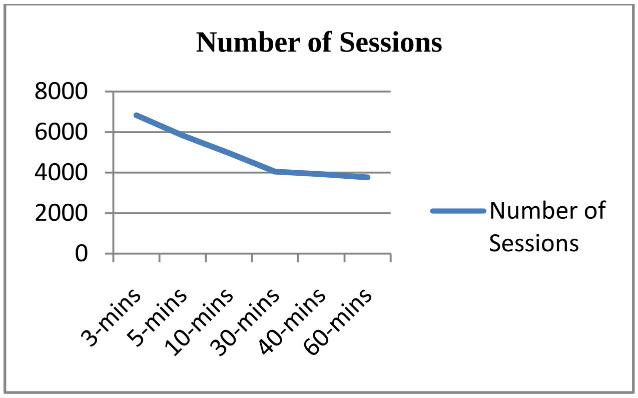

Fig. 3: Number of Sessions

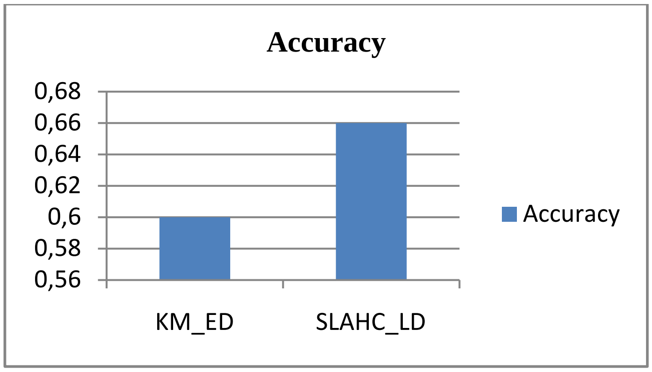

Fig. 4: Accuracy Comparison between KM_ED and SLAHC_LD

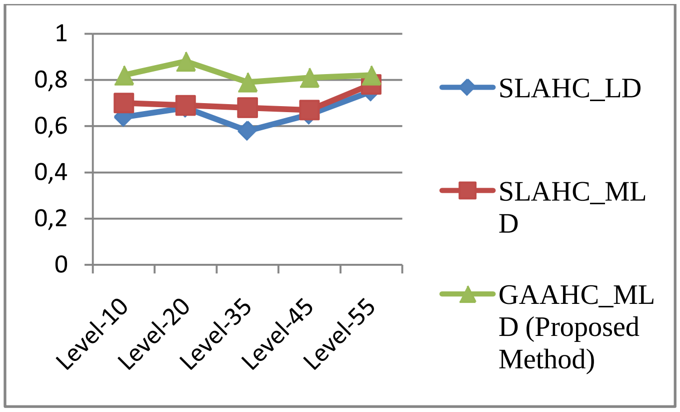

Fig. 5: Comparison of Accuracy between SLAHC_LD, SLAHC_MLD and GAAHC_MLD

7 Conclusion

In the proposed work personalized recommendation of the web pages to the web users with an average accuracy of 0.82 is achieved, assuming the session timeout to 30mins with minimum of five page views in a session. The proposed method based on Group Average Agglomerative Hierarchical Clustering using Modified Levenshtein Distance and Page Ranking using Access Frequency can be used to personalize the web user's home page with the links of the web pages the user may likely to visit. The proposed is compared with other existing methodologies like KM_ED, SLAHC_LD and SLAHC_MLD and has achieved an average accuracy of 0.82. Hence, Group Average Agglomerative Hierarchical Clustering using Modified Levenshtein Distance is an important discovery made in the proposed work.

7.1 Future Directions

The proposed work focuses on the sequence of page visits made by the user to generate the personalized recommendation to the web users. In future, the present work can be enhanced by using the Semantic Group Average Agglomerative Hierarchical Clustering using the metric that computes the semantic similarity between the web page title tag string and web user profile instead of using Modified Levenshtein Distance.

References

[1] Hiral Y. Modi and Meera Narvekar, “Enhancement of Online Web Recommendation System Using a Hybrid Clustering and Pattern Matching Approach”, *International Conference on Nascent Technologies in the Engineering Field (ICNTE-2015)*

[2] A Vinupriya and S. Gomathi, “Web Page Personalization and Link Prediction Using Generalized Inverted Index and Flame Clustering”, *International Conference on Computer Communication and Informatics (ICCCI), Jan-07-09, 2016, Coimbatore, India.*

[3] Mercy Paul Selvan, A. Chandrashekar, Deepak R Babu and A. Krishna Teja, “Efficient Ranking Based on Web Page Importance and Personalized Search”, *International Conference on Communications and Signal Processing (ICCSP), IEEE, 2015.*

[4] Zhongyun Ying, Zhorong Zhou and Goufeng Zhu, “Research on Personalized Web Page Recommendation Algorithm Based on User Context and Collaborative Filtering”, *Fourth IEEE International Conference on Software Engineering and Service Science (ICSESS),* IEEE, 30th Sep 2013, Beijing, China.

[5] Gerrad Deepak, J Sheeba Priyadarshini and M S Hareesh Babu, “A Differential Semantic Algorithm for Query Relevant Web Page Recommendation”, *IEEE International Conference on Advances in Computer Applications (ICACA),* 24th Oct 2016, Coimbatore, India.

[6] Korinna Bade and Andreas Nurberger, “Personalized Hierarchical Clustering”, *IEEE/WIC/ACM International Conference on Web Intelligence,* Hong Kong, China, 2007

[7] Dipa Dixit and Jayant Gadge, “Automatic Recommendation for Online Users Using Web Usage Mining”, *International Journal of Managing Information Technology (IJMT),* Vol 2, Issue 3, August-2010.

[8] K Suneetha and M. Usha Rani, “Performance Analysis of Web Page Recommendation Algorithm Based on Weighted Sequential Patterns and Markov Model”, *International Journal of Computer Science Issues,* Vol 10, Issue 1, No. 3, Jan-2013, ISSN (Print): 1694-0784, ISSN (Online): 1694-0814.

[9] Neeraj Iyer, Alex Dcunha, Akshay Desai and Kavita Jain, “Survey on Online Recommendation Using Web Usage Mining”, *International Journal of Computer Science and Information Technologies,* vol 6, Issue 2, 2015, ISSN: 1465-1467.

[10] V. Chitraa and Antony Selvadoss Thanamani, “Recommendation of Web Pages for Online Users Using Web Log Data”, *International Journal of Science and Research (IJSR),*

[11] Zohreh Anari and Babak Anari “Determining the Similarity of Web Pages Based on Learning Automata and Probabilistic Grammar”, *Advances in Computer Science an International Journal (ACSIJ),* Vol 4, Issue 3, Page No. 15, May-2015, ISSN: 2322-5157.

[12] S. Ramanamurthy and G. Anuradha, “Implementation of Page Ranking Using Genetic Algorithm”, *Proceedings of Seventh IRF International Conference,* 12th Oct 2014, Goa, India, ISBN: 978-93-84209-57-5.

[13] Harish Kumar B T, Vibha L and Venugopal K R, “Web Page Access Prediction Using pal K R, “Web Page Access Prediction Using Hierarchical Clustering Based on Modified Levenshtein Distance and Higher Order Markov Model”, *IEEE International Conference organized by IEEE Region 10 Symposium (TENSYMP-2016),* Bali, Indonesia on 9th to 11th May 2016.