International Journal of Soft Computing 8 (6): 381-385, 2013
ISSN: 1816-9503

User Interest Estimation Using Behavior Monitoring Measure

[1]R. Krishnamoorthy and [2]K.R. Suneetha
[1]Department of Computer Science and Engineering, University College of Engineering,
A Constituent College of Anna University, BIT Campus,
Chennai, Tiruchirappalli, Tamil Nadu, India
[2]Department of Computer Science and Engineering, Bangalore Institute of Technology,
V.V. Puram, K.R. Road, Bengaluru, 560004 Karnataka, India

Abstract: User interest estimation is an imperative part of any personalized information retrieval system whose quality is decided based on the accuracy of user interest description. This study proposes a Behavior Monitoring Measurement System which is an advanced metric model for estimating the degree of user's interest based on his browsing behaviors and histories. Since, browsing behaviors are the valuable indicators of user's interest, this metric considers browsing behaviors such as duration of time spent in browsing each page and various other user activities. As the user behavior changes the measurement values would get updated and finally degree of satisfaction would be estimated. The calculation for the degree of the matic interest is based on Clustering Rule Method. The mathematical formulation for calculating the interested degree is also presented in this study. The experimental results confirm that the proposed system reflects user's satisfaction degree accurately and dynamically.

Key words: Personalized Information Retrieval System, behavior monitoring measure, clustering rule, degree, metric

INTRODUCTION

The advent of the World Wide Web and information updating technologies has made the analysis and discovery of required information snarled (Srivastava *et al*., 2002; Kosala and Blockeel, 2000; (Cooley *et al*., 1999). Inferring the user interests has become a great challenge for most of the web personalized recommendation systems.

Estimation of user interests relies greatly on the browsing behaviors and history which includes his time duration on each page and various other user activities (Suneetha and Krishnamoorti, 2009; Li *et al*., 2008). For monitoring these activities an advanced metric model called behavior monitoring metric is proposed in this study. This metric model uses a computer software program which acts as a collector as well as an estimator. It tracks the browsing activities of a user and collect his browsing behaviors using which it estimates his interests. Once set up, the behavior rating would be updated for every user behavior changes for the requested web page and finally the user's satisfaction degree would be estimated. Followed by this the degree of thematic interest is estimated using the selected pages which has high satisfaction degree. This is carried out based on the clustering analysis.

LITERATURE REVIEW

Here, researchers review some of the methods available in the literature. In recent years, a large amount of work has been conducted in the area of web mining. Many studies have been performed on web usage mining (Nasraoui *et al*., 2008; Li and Zhong, 2006) understand web user activities. A number of studies have concentrated on applying data mining techniques identify the behavior of frequent users. To improve web site retention the decision tree C4.5 algorithm was used by Pabarskaite (2003) in which they perform classification to carry out web user behavior analysis.

Page contents are important for finding user interests. According to Li *et al*. (2008) the clustering algorithm first utilizes Kaufman Approach (KA) for initialization which initializes clustering by successively selecting representative page instances until m instances have been found. Then, it uses the selected m seeds as

Corresponding Author: K.R. Suneetha, Department of Computer Science and Engineering, Bangalore Institute of Technology, V.V. Puram, K.R. Road, Bengaluru, 560004 Karnataka, India

the initial centroids and finally performs the spherical k-means algorithm (Qinghong *et al*., 2012) to divide all the pages in to m clusters. Based on the clustering results, some keywords are extracted to represent user interests and the summarization method is implemented to provide some detailed information for each interest.

User interest model is the key part of the personal service system and the accuracy of user interest description directly decides the quality of the personal service system. This study collects information including browsing behavior from the web users register information, server diaries as well as BHO components. After the preprocessing of the information collected and calculated user interest rate, the study will analyze and process the data and then build a user interest model.

USER BEHAVIORS COLLECTION

The estimation of user interest is based on the combination of user's browsing behaviors. This browsing behavior is classified in to two categories, namely probative behavior and periphrastic behavior. Save page, print page and number of visits to the same page are the user activities which come under probative behavior category. Time spent on a page, mouse, keyboard and scrollbar activities, etc. are the user activities which come under periphrastic behavior category.

Probative behavior: With the analysis of user's probative behaviors, save page and print page activities draws us to assume that the user is very interested on this page. Saving a page to the bookmark signifies that the user is really interested in these pages and will be invoked each time the user revisits it (Zheng *et al*., 2010).

Periphrastic behavior: Considering user's browsing time on a page, the longer the dwelling time, the more is the user interested on the page. Therefore, the total time spent on a web page is the best indicator of the interest degree. The scrollbar and mouse activities are measured in terms of the number of mouse clicks and time spent in scrolling and moving the mouse over a particular web page. As far as the keyboard activities are concerned, they are measured in terms of the number of key hits which are page up, page down, up arrow and down arrow (Claypool *et al*., 2001).

USER BEHAVIOR RATE ANALYSIS USING THE BEHAVIOR MONITORING SYSTEM

User interest rate analysis based on browsing behavior: Once the system starts monitoring the user who has logged in to the internet, it would keep a track of the user from the time he signs in till the time the user signs out.

During this time, the system would also keep track of all the requested web pages and the user browsing behaviors.

Once the requested web page is rendered to the user, the system would keep a record of the time spent on a page from the loading time until the users exit. During this period various probative and periphrastic browsing behaviors of the user are captured and the degree of interest for each activity is estimated. The estimated value signifies the satisfaction value which is entered in appropriate scoring option box (Qinghong *et al*., 2012). The satisfaction values are divided into three grades: uninterested, interested and very interested where the value for each grade would be 0, 1 and values >1. The behavior monitoring system estimates the user activities using web page dwelling time, mouse/keyboard and scrollbar activities.

Web page dwelling time: Since, the total time spent on a web page reliably indicates the interest degree of the user, a graph is drawn on the user interest rate against the total time spent on the page for analysis purpose. Let Ta be the total time spent on a page. If the page is analyzed for a time period greater than T/N the page is considered to be very interesting. If it is analyzed for a period of T/N, researchers conclude that the user is interested in a page. If it is analyzed for a time period lesser than T/N, it signifies that the user is uninterested in that page.

Mouse/keyboard and scrollbar activities: While analyzing the web page, this system also measured the time spent for mouse/keyboard activities and time utilized scrollbar activities. Assume that the time spent on a particular web page is T_a and the time spent for mouse/keyboard activities and scrollbar activities to be T_{mk} and T_s.

If the time spent in mouse/keyboard activities and scrollbar activities are lesser T_a/T_{mk} and T_a/T_s, the page would be uninteresting. If it is more than T_a/T_{mk} and T_a/T_s then it is very interesting. If the time spent is exactly T_a/T_{mk} and T_a/T_s then the page is considered to be interesting. Through these statistics, the average satisfaction degree for each page would be calculated. This is formulated in Table 1.

Table 1: User Interest rate analysis based on browsing behavior

Probative and periphrastic browsing behaviors	Degree of satisfaction		
	Uninterested	Interested	Very interested
Page X			
Req_Id	0	1	>1
Save page	0	1	>1
Print page	0	1	>1
Browsing time	<T/N	T/N	>T/N
Mouse/Keyboard activities	$<T_a/T_{mk}$	T_a/T_{mk}	$>T_a/T_{mk}$
Scrollbar activities	$<T_a/T_s$	T_a/T_s	$>T_a/T_s$
Average degree of satisfaction	<0.5	0.5	>0.9

Table 2: User interest degree estimation is done for all the tracked pages

	Probative browsing behavior		Periphrastic browsing behavior			
Pages	Save page	Print page	Browsing time (msec)	Mouse/keyboard activities (msec)	Scrollbar activities (msec)	Interest degree estimation
A-56	1	0	50,000	9,000	10,000	Interested
B-98	0	0	60,000	9,000	10,000	Uninterested
A-56-1	1	1	60,000	10,000	20,000	Very interested
B-34	1	0	50,000	10,000	15,000	Interested
A-56-2	1	1	40,000	9,000	20,000	Interested
A-84	0	1	80,000	12,000	25,000	Interested
A-90	0	1	60,000	14,000	22,000	Interested
X-45	0	1	60,000	12,000	20,000	Interested
A-84-1	1	0	70,000	11,000	30,000	Interested
B-34-1	1	0	60,000	10,000	15,000	Interested

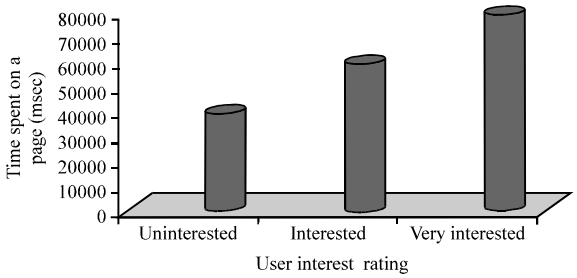

Fig. 1: Time spent on a page vs. user interest rating

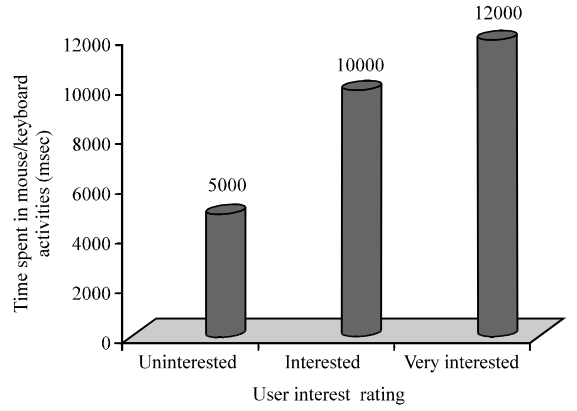

Fig. 2: Time spent in mouse/keyboard activities vs. user interest rating

Now, researchers would start running the test by assuming that the user assesses the internet for a time interval T which is taken as 6,00,000 msec (about 10 min) and tracked web pages be N which is taken as 10. Thus, according to the formulation depicted in Table 1, the time spent on a page, mouse/keyboard and scrollbar activities are calculated for the given data values and depicted graphically in Fig. 1-3. Finally, the user interest degree estimation is carried out for all the tracked pages and portrayed in Table 2.

Theme interest degree calculation based on clustering rule: After obtaining the users interest degree for the

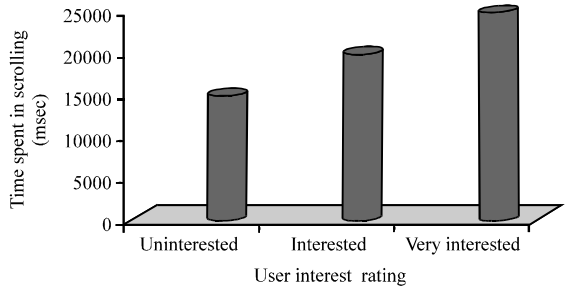

Fig. 3: Time spent in scrolling vs. user interest rating

given pages, computing the time spent on web sited theme interest degree is not sufficient to describe the user's interest for the given theme. Now, researchers find the correlation between all the pages whose degree of interest is rated as interesting and very interesting by applying rules stated in Table 2. In order to figure out the degree of correlation between selected pages for the corresponding theme researchers draw a spanning tree for the selected pages. The spanning tree method is an agglomerative hierarchical clustering technique whose main motto is to partition the selected pages in to groups. Thus, the members of a group are as similar as possible and different groups are as dissimilar as possible.

Thus, researchers start running the test by considering five pages whose average satisfaction degree is marked to be very interesting out of the ten tracked pages. The probative and periphrastic browsing behavior counts for each of these pages would be obtained from the meter. These parameters would serve as the data for clustering.

Assume that the counts for the probative and periphrastic behaviors for each of the pages as mentioned in Table 3 as follows. The approach starts with n clusters of one page each and successively joins the nearest clusters. The steps for this process are illustrated:

- In step 1 consider each page to be cluster in which case there would be five clusters

Table 3: Probative and periphrastic behaviors count

Pages	Probative behavior count	Periphrastic behavior count
A-56	2	4
A-84	3	5
A-90	1	6
X-45	4	3
B-34	5	2

Table 4: Euclidian distance measure

Pages (DM)	Pages (EuDM) A-56	A-84	A-90	X-45	B-34
A-56	0	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{13}$
A-84		0	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{13}$
A-90			0	$\sqrt{18}$	$\sqrt{32}$
X-45				0	$\sqrt{2}$
B-34					0

Table 5: Distance matrix for three clusters

Pages (DM)	Pages (EuDM) [A-56] [A-84]	A-90	[X-45] [B-34]
[A-56] [A-84]	0	$\sqrt{4.5}$	$\sqrt{8}$
A-90		0	$\sqrt{24.5}$
[X-45] [B-34]			0

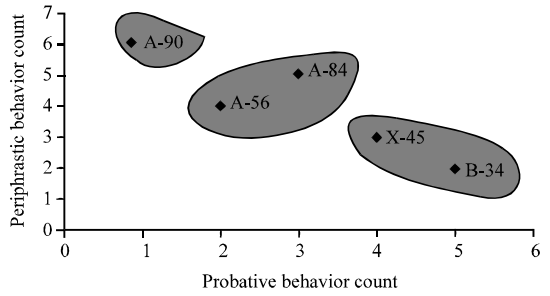

Fig. 4: Clustered points

- In step 2 the parameter's count for each of the pages are considered as centroids which are {2, 4}, {3, 5}, {1, 6}, {4, 3} and {5, 2}. These are depicted as points in Fig. 4
- In step 3 using the Euclidean distance measure, the distance matrix for the pages are measured in Table 4 as follows
- In step 4 the minimum intercluster distance is found to be 2 between the pages A-56 and A-84 and between $\sqrt{X\text{-}45}$ and B-34. This signifies the degree of correlation between the two pairs which are finally merged
- In step 5 the centroids of the cluster pair [A-56] [A-84] is {(2+3)/2, (4+5)/2} which would be {2.5, 4.5}. Similarly the centroids of the pair [X-45] [B-34] would be {4.5, 2.5}. Thus, there would be three clusters now and the distance matrix for it is portrayed below (Table 5)
- In step 6 the minimum intercluster distance [A-56] [A-84] and A-90 is 4.5. Thus, these two clusters are merged and its centroids are found to be $\sqrt{\{2, 5\}}$ for which the distance matrix would be (Table 6)

Table 6: Distance matrix for next two clusters

Pages (DM)	Pages (EuDM) [A-56] [A-84] [A-90]	[X-45] [B-34]
[A-56] [A-84] [A-90]	0	$\sqrt{12.5}$
[X-45] [B-34]	-	0

- In step 7 thus, the minimum intercluster distance between the pair [A-56] [A-84] [A-90] and [X-45] [B-34] is found to be $\sqrt{12.5}$. This signifies the degree of correlation between the selected pages for the given user theme. Now the merger of these two pairs would produce single cluster [A-56] [A-84] [A-90] [X-45] [B-34]

CONCLUSION

In this study, researchers have proposed an advanced software system for behavior monitoring based on the degree of user's interest on his browsing behaviors and histories. The degree of thematic interest is estimated using clustering analysis method. The adopted methods are proved to be accurate and effective experimentally with good comprehensive considerations. Thus, the experimental results confirm the degree of user interest accurately with high efficiency and accuracy. Future researches in this direction can be inclusion of a fuzzy logic based approach to measure the user's interest qualitatively.

REFERENCES

Claypool, M., D. Brown, P. Le and M. Waseda, 2001. Inferring user interest. IEEE Internet Comput., 5: 32-39.

Cooley, R., B. Mobasher and J. Srivastava, 1999. Data preparation for mining world wide web browsing. J. Knowledge Inform. Syst., 1: 5-32.

Kosala, R. and H. Blockeel, 2000. Web mining research: A survey. ACM SIGKDD Explorat. Newslett., 2: 1-15.

Li, F., Y. Li, Y. Wu, K. Zhou, F. Li, X. Wang and B. Liu, 2008. Combining browsing behaviors and page contents for finding user interests. Proceedings of the 8th International Workshop on Autonomous Systems-Self-Organization, Management and Control, October 6-7, 2008, Shanghai Jiao Tong University, Shanghai, China, pp: 149-156.

Li, Y. and N. Zhong, 2006. Mining ontology for automatically acquiring web user information needs. IEEE Trans. Knowl. Data Eng., 18: 554-568.

Nasraoui, O., M. Soliman, E. Saka, A. Badia and R. Germain, 2008. A web usage mining framework for mining evolving user profile in dynamic web sites. IEEE Trans. Knowl. Data Eng., 20: 202-215.

Pabarskaite, Z., 2003. Decision trees for web log mining. Intell. Data Anal., 7: 141-154.

Qinghong, Y., H. Hao and X. Neng, 2012. The research on user interest model based on quantization browsing behavior. Proceeding of the 7th International Conference on Computer Science and Education, July 14-17, 2012, Melbourne, Australia.

Srivastava, J., P. Desikan and V. Kumar, 2002. Web-Mining: Accomplishments and future directions. Technical Report Computer Science Department, University of Minnesota, Minneapolis, pp: 51-61.

Suneetha, K.R. and R. Krishnamoorti, 2009. Identifying user behavior by analyzing web server access log file. Int. J. Comput. Sci. Network Secur., 9: 327-332.

Zheng, L. S. Cui, D. Yue and X. Zhao, 2010. User interest modeling based on browsing behavior. Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering, August 20-22, 2010, Phuket, Thailand.