Assessment and Validating the Quality of Educational Web Sites using Subtractive Clustering

Ramin Afshoon
Department of Computer Engineering, College of Computer, Khouzestan Science and Research Branch, Islamic Azad University, ahvaz, Iran

Ali Harounabadi
Department of Computer Engineering, Tehran Center Branch, Islamic Azad University, Tehran, Iran

Javad MirAbedini
Department of Computer Engineering, Tehran Center Branch, Islamic Azad University, Tehran, Iran

ABSTRACT
Researchers have studied qualitative and quantitative methods to assess the quality of website. Previous studies had determined criteria such as quality of service. Human behavior, namely the objective perspective, is the essential source to obtain human thinking and real doings. For this reason, data mining approaches are used to acquire the objective source. In this research, proposed subtractive clustering is applied in evaluating educational web sites from the fuzzy objective perspective. An empirical study is carried out to validate the model capability. Results indicate that in the recommended algorithm are closer to the real data.

Keywords
Web site quality, Data mining, Subtractive clustering

1. INTRODUCTION
Nowadays with the advancement and development in information and network technologies, lots of data have been digitized to reveal information for users by the construction of Web sites. Since Web sites serve as a major portal to connect with these information, evaluating the quality or utility of Web sites becomes necessarily as a way to understand whether users are satisfied with these data or not. Previous studies had determined criteria such as quality of service for evaluating web sites [1].

Web mining is, using data mining tools to discover knowledge from various sources in the web and according to sources type that mined, to be classified in various fields of research. One of the important tools in web mining is mining of web user's behavior that is considered as away to discover the potential knowledge of web user's interaction. Data mining approaches are used to acquire the objective source. Based on the data gathered from the World Wide Web, the web mining is divided into 3 groups: web content mining, web structure mining, and web usage mining [2]. The web content mining deals with information or the science discovered from web pages content. The web structure mining discovers the relations between web pages, by analyzing the web structure. According to the hyperlinks, the web structure mining organizes the web pages in groups and produces relative patterns, such as similarities and relations between various web sites. The web usage mining is the process of discovering the fact that what are the users searching for in the internet.

Human behavior, namely the objective perspective, is the other essential source to obtain human thinking and real doings. For the Web behavior, the past browsing logs are used from users to analyze their actual usages of web sites. As the result, if a user visits a Web site with a longer time and clicks Web pages more times, it shows that she/he is eager with the presentation and content of the Web site. Therefore, the user's certain perception is gathered according to a series of the real behaviors without interrupting the user.

The rest of this paper is organized as follows:
In the second section the research background has been described. In the third part the previous methods will be investigated and in the fourth, the suggested method of this study is to be introduced. The fifth section explains the details regarding the implementation and analysis of the suggested method and eventually on the sixth part, the result and the strong points of the suggested method are going to be explained.

2. BACKGROUNDS
In this section the subtractive clustering used in this article is explained.

2.1 Clustering
Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) [3]. Clustering of numerical data forms the basis of many classification and system modeling algorithms. The purpose of clustering is to distill natural grouping of data form a large data set, producing a concise representation of a system's behavior.

2.2 Subtractive Clustering
Consider a collection of n data points $\{x_1, x_2, \ldots, x_n\}$ in an M-dimensional space. Without loss of generality, it is assumed that the data points have been normalized in each dimension so that their coordinate ranges in each dimension are equal, i.e., the data points are bounded by a hypercube. Each data point is considered as a potential cluster center and defines a measure of the potential of data point x_i as

$$P_i = \sum_{i=0}^{n} e^{-\propto \|x_i - x_j\|^2} \quad (1)$$

Where

$$\propto = {}^{4}/_{r_a{}^2} \quad (2)$$

and r_a is a positive constant. Thus, the measure of potential for a data point is a function of its distances to all other data points [4]. A data point with many neighboring data points will have a high potential value. The constant r_a is effectively the radius defining a neighborhood; data points outside this radius have little influence on the potential. This measure of potential differs from that proposed by Yager and Filev in two ways: (1) the potential is associated with an actual data point instead of a

grid point; (2) the influence of a neighboring data points decays exponentially with the square of the distance instead of the distance itself.

After the potential of every data point has been computed, the data point is selected with the highest potential as the first cluster center. Let x_1^* be the location of the first cluster center and p_1^* be its potential value. Then the potential of each data point x_i is revised by the formula

$$P_i=P_i-P_1^* e^{-\beta\|x_i-x_1^*\|^2} \quad (3)$$

Where

$$\beta=4/r_b^2 \quad (4)$$

and r_b is a positive constant. Thus, an amount of potential is subtracted from each data point as a function of its distance from the first cluster center. To avoid obtaining closely spaced cluster centers, r_b is set to be somewhat greater than r_a; a good choice is $r_b=1.5\ r_a$.

When the potential of all data points have been revised according to eq. (3), the data point with the highest remaining potential is selected as the second cluster center.

Each cluster center x_1^* is considered as a fuzzy rule that describes the system behavior. Given an input vector y, the degree to which rule I is fulfilled is defined as

$$\mu_i=e^{-\propto\|y-x_i^*\|^2} \quad (5)$$

A comparison study has been performed between FCM clustering algorithm and subtractive clustering algorithm according to their capabilities to model a set of non-linear systems and experimental data. The comparison is based on validity measurement of their clustering results. The number of clusters is changed for the fuzzy c-mean algorithm. The validity results are calculated for several cases. As for subtractive clustering, the radii parameter is changed to obtain different number of clusters. Generally, increasing the number of generated cluster yields an improvement in the validity index value. The optimal modeling results are obtained when the validity indices are on their optimal values. Also, the models generated from subtractive clustering usually are more accurate than those generated using FCM algorithm. A training algorithm is needed to accurately generate models using FCM. However, subtractive clustering does not need training algorithm. FCM has inconsistency problem where different runs of the FCM yields different results. On the other hand, subtractive algorithm produces consistent results [5].

3. RELATED WORKS

Guo and et al used fuzzy clustering to customer relationship management of the securities industry [6].

Huang and Huang proposed an integrated decision model applied in evaluating educational Web sites from the fuzzy subjective and objective perspectives. The former source is extracted by inquiring human opinion using a questionnaire, while the latter is gained automatically by a data mining technique, fuzzy clustering [7].

Pamutha and et al focused on the preprocessing of the web log file methods that can be used for the task of session identification from web log file. The work also produced statistical information of user session, such as: (1) total unique IPs; (2) total unique pages; (3) total sessions; (4) Session length and (5) the frequency visited pages. After preprocessing completed, the result will be used for mining user access patterns [8].

Santra and Jayasudha studied the behavior of the interested users instead of spending time in overall behavior. The existing model used enhanced version of decision tree algorithm C4.5. They proposed to use the Naive Bayesian Classification algorithm for classifying the interested users and presented a comparison study of using enhanced version of decision tree algorithm C4.5 and Naive Bayesian Classification algorithm for identifying interested users. The performance of this algorithm is measured for web log data with session based timing, page visits, repeated user profiling, and page depth to the site length. Experimental results conducted shows that the performance metric i.e., time taken and memory to classify the web log files are more efficient when compared to existing C4.5 algorithm [9].

Demirli and et al proposed an extended subtractive clustering based fuzzy system identification method and the Sugeno type reasoning mechanism are used for modeling job sequencing problems. This approach can be used to build a fuzzy model of the sequencing system from an existing sequence (output data) and possible job attributes (input data). The single machine weighted flow time problem is used as an example to demonstrate the proposed methodology. The effects of data scarcity on the modeling performance are studied by using three data sets with varying degrees of available data. Furthermore, a parametric search on various clustering parameters is performed to identify the best model. As a result of parametric search, ranges of clustering parameters that provide best models are also identified [10].

4. THE PROPOSED METHOD

The proposed method is used web usage mining and subtractive clustering for evaluating service quality of educational web sites. In this study the log file of HEEACT web server was utilized. Since there are unprocessed data in the log files, they must be subjected to some preprocessing in order to be used in web mining, which in this article, data cleansing, distinguishing the users from each other and identifying sessions per user is considered [11]. Its flowchart is shown in Fig 1.

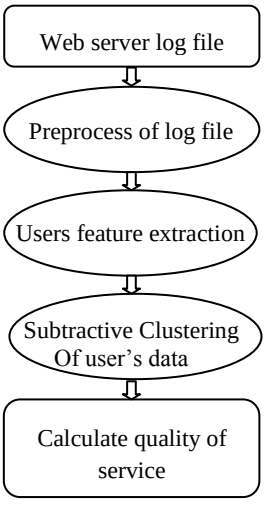

Fig 1: Evaluating service quality of educational web sites

The objective data is set out to collect from the user's browsing behavior in Web sites. Different kinds of statistical analyses can be used to extract knowledge about users to visit a Web site, such as viewing time (VT), page view frequency (PVF), or

length of a navigational path (LNP) [11]. In this experiment, the first 2 parameters are gathered, because LNP is only used to find Web usage patterns in Web mining. Then, the techniques of data cleaning, user identification and session identification are employed for the data preprocessing.

Next, the subtractive clustering approach is executed to gain the Web sites' vectors and grades from the two variables, VT and PVF, by the log file users. The averages of the two variables are computed as the clustering attributes.

5. EVALUATION AND IMPLEMENTATION OF THE PROPOSED METHOD

In this section, the details of implementation of the suggested method are explained. To implement the components of the suggested system, MATLAB and Microsoft SQL Server 2008 software were used. To evaluate the proposed method, HEEACT's log file was used. For controlling the educational quality in Taiwan, a series of university evaluations for every five years have been handled by the Ministry of Education [12].

A group of professionals, called Higher Education Evaluation & Accreditation Council of Taiwan (HEEACT), are commissioned to develop standards and evaluate higher education institutions. Departments or graduate institutes in a university has to be evaluated by five dimensions: (1) their objective, feature, and self-improvement, (2) course design and teaching, (3) student learning and affair, (4) the production of the academic researches, and (5) the achievement of the graduated students[12].

Unlike the traditional education resources presenting in paper, all digital education resources are stored in databases and shown in Web sites. Nowadays, students have changed their learning behavior into accessing them by Internet.

According to the evaluation report announced by HEEACT, twenty-four departments were randomly sampled from forty-five evaluated universities or colleges in Taiwan, and their portal Web sites were as our studying subjects. The field distributions involved education (Edu), engineering (Eng), humanities (Hum), and management (Man). The twenty-four departments have been evaluated by HEEACT and have announced the assessment results. Their fields and Web URLs are listed in Table 1.

Table 1. Data of the twenty-four departments

No.	Field	URL
1	Eng	http://www.cs.nthu.edu.tw/
2	Hum	http://www.nchu.edu.tw/foreign/
3	Man	http://www.im.ncue.edu.tw/
4	Eng	http://www.cse.yzu.edu.tw/
5	Edu	http://www.ntnu.edu.tw/spe/news.html
6	Edu	http://dpts.nttu.edu.tw/sped/contents/news/news_list.asp?menuID¼41
7	Man	http://dept.hku.edu.tw/mis/
8	Man	http://ibd.ndhu.edu.tw/main.php
9	Hum	http://fll.hcu.edu.tw/front/bin/home.phtml
10	Edu	http://social.tmue.edu.tw/front/bin/home.phtml
11	Hum	http://doflal.niu.edu.tw/news/news.php?class¼4101
12	Man	http://www.im.knu.edu.tw/cht/main.asp
13	Edu	http://www.ntcu.edu.tw/sse/webweb/index2.html
14	Man	http://www.iba.leader.edu.tw/
15	Man	http://www.dwu.edu.tw/winformation/mis95/index-2.htm
16	Man	http://ib.toko.edu.tw/newsite/index.asp
17	Edu	http://social.ntue.edu.tw/home.htm
18	Edu	http://spec.tmue.edu.tw/front/bin/home.phtml
19	Eng	http://www.ncyu.edu.tw/csie/
20	Hum	http://www.fl.chu.edu.tw/news.htm
21	Edu	http://dpts.nttu.edu.tw/soc/contents/news/news_list.asp?menuID¼4285
22	Edu	http://163.23.205.45/sped/
23	Man	http://dept.hku.edu.tw/iba/index1.htm
24	Eng	http://www.csie.ndhu.edu.tw/webv3/cht/?board¼news&main¼news_bd01

The descriptive statistics of viewing time (VT), page view frequency (PVF) are presented for each Web site in Table 2.

Table 2. Descriptive statistics

No.	VT	PVF	No.	VT	PVF
1	121.9	2.43	13	51.75	4.25
2	28.67	2	14	27.6	3
3	81.75	6.38	15	75.6	5
4	66	3.82	16	38.67	3.33
5	89.33	3.67	17	49.6	3
6	44.5	1.67	18	51.5	2.5
7	93.52	5.25	19	129.3	1.5
8	62.33	1.67	20	41.5	3.5
9	69.5	4.8	21	46.67	2.17
10	64	2.86	22	68.25	4.25
11	103.28	30.25	23	30	8
12	54.1	2.2	24	66.6	7.6

To gain a proper result in proposed method, accept and reject ratios are increased from 0.01 to 0.99 with 0.001 step size. In this method, utilizing a point to point similarity, the best number of clusters is determined automatically [13]. The subtractive clustering algorithm is applied on the described data set. The algorithm is developed in MATLAB [14] and the results are shown in Fig 2.

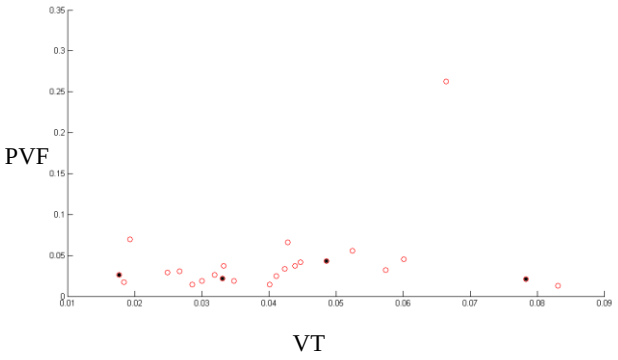

Fig 2: Cluster centers in MATLAB

By setting clustering values, four clusters as the optimal number of clusters obtained. To classify the quality of Web sites are chosen into four grades, Excellent, Good, Middle and Poor. To

judge the four grades with respect to two variables, it is specified that the determinant of Excellent is long VT and high PVF, Good is long VT and low PVF, Middle is middle VT and middle PVF, Poor is short VT and low PVF. Results are compared with FuzzSOP [15] in Table 3.

Table 3. Evaluation of service quality criteria

No.	FuzzSOP	Proposed method
1	Excellent	Excellent
7	Good	Good
8	Middle	Middle
15	Good	Middle
20	Poor	Poor

The results are very similar to the FuzzSOP method. Subtractive clustering method is superior to the fuzzy clustering.

6. CONCLUSION

Evaluating the quality or utility of Web sites becomes necessarily as a way to understand whether users are satisfied with these data or not. Previous studies had determined criteria such as quality of service for evaluating web sites.

In this article, a subtractive clustering algorithm is presented for evaluating service quality of educational web sites. Subtractive clustering method is superior to the fuzzy clustering. So it can be used as a reference method for assessing the service quality of educational Web sites.

7. REFERENCES

[1] Lin, H.F, "Measuring online learning systems success: Applying the updated DeLone and McLean's model", Cyber Psychology and Behavior, Vol. 10, Issue. 6, pp. 817–820, 2007.

[2] Mustapasa, O., Karahoca, D., Karahoca, A., Yucel, A., Uzunboylu, H. 2010. Implementation of semantic web mining on e-learning. In: proc. of Social and Behavioral Sciences, vol.2, Issue 2, pp. 5820-5823.

[3] Jain, A.K., Murty, M.N., Flynn, P.J. 1999. Data clustering: a review. ACM Computing Surveys (CSUR), Vol.31, Issue 3, pp. 264-323.

[4] Chiu, S.L. 1994. Fuzzy model identification based on cluster estimation. Journal of Intelligent and Fuzzy Systems, Vol. 2, pp. 267-278.

[5] Bataineh, K.M, Naji, M, Seqar, M. 2011. A comparison study between various fuzzy clustering algorithms. Jordan Journal of Mechanical and Industrial Engineering, Vol. 5, No. 4, pp. 335-343.

[6] Guo, L., Zhang, M., Sun, L., Wang, Z. 2006. Fuzzy clustering model of CRM in securities trade. Proceedings of the 6thWorld Congress on Intelligent Control and Automation (WCICA), pp. 6052-6054.

[7] Huang, T.C, Huang, C. 2010. An integrated decision model for evaluating educational web sites from the fuzzy subjective and objective perspectives. Computers & Education. Elsevier, Vol. 55, pp. 616-629.

[8] Pamutha, T., Chimphlee, S., Kimpan, C., Sanguansat, P. 2012. "Data preprocessing on web server log files for mining users access patterns". International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol. 2, No. 2, pp. 92-98.

[9] Santra, A.K., Jayasudha, S. 2012. Classification of web log data to identify interested users using naïve Bayesian classification. International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, pp. 381-387.

[10] Demirli, K., Cheng, S.X., Muthukumaran, P. 2003. Subtractive clustering based modeling of job sequencing with parametric search. Fuzzy Sets and Systems, Vol. 137, Issue 2, pp. 235–270.

[11] Cooley, R., Mobasher, R., Srivastava, J. 1999. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, vol.1, pp.5–32.

[12] Heeact. 2009. Higher education evaluation & accreditation council of Taiwan. [WWW document]. Available from. http://www.heeact.edu.tw/mp.asp?mp¼4

[13] Sahebi S., Oroumchian F. and Khosravi R., 2008, an enhanced similarity measure for utilizing site structure in web personalization systems, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, IEEExplore, pp. 82-85.

[14] MATLAB Software. http://www.mathworks.com.

[15] Huang, T.C, Huang, C. 2010. An integrated decision model for evaluating educational web sites from the fuzzy subjective and objective perspectives. Computers & Education. Elsevier, Vol. 55, pp. 616-629.