International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
42 – 44

Combining Content Mining and Usage Mining for Web Data Extraction (CCMUM)

Neeraj Raheja[1], Dr.Vijay Kumar Katiyar[2]

Associate Professor1, Professor2
[1, 2] Department of Computer Science and Engineering,
M.M.University, Mullana (Ambala), Haryana, India
neeraj_raheja2003@mmumullana.org, katiyarvk@mmumullana.org[2]

Abstract- The demand of Web data extraction is increasing due to large amount of data available over the Websites or Web Sources. This research work provides a model CCMUM which combines Web content mining and Web usage mining. Thus, CCMUM acts as a complete Web wrapper which uses the advantage of both content mining as well as usage mining. The results of CCMUM will be compared with the N-gram approch. CCMUM will provide more efficient results as compared to N-gram.

Keywords-*Content Mining, Usage Mining, Combining Content Mining and Usage Mining*

*****

I INTRODUCTION

The data available over the Websites is increasing at a tremendous rate due to the demand of the users or market. Thus, data provided over the Websites must fulfill both criterions i.e. firs t it must include large amount of data and second it must consider the user perspective or user choice for popularity of Website. To extract data according to first criterion, Web content mining comes into picture and to extract data according to second criterion Web usage mining comes into scenerio [1][3].

Web content mining is used to extract content from large amount of data and Web usage mining is used to record user behavior in the weblogs [4][6]. Combination of both types of mining provides efficient results in terms of quality, accuracy and efficiency and also improves the popularity of the Website[5]. This process also helps in search engine optimazation and improves the ranking of the Website accordingly[7][8][9]. Another advantage of this process is that webmasters can improve the design and organization of websites based on results obtained [1] [2].

II LITERATRE REVIEW

Haibin Liu et.al (1) proposed an approach for automatic Web data extraction which uses integration of usage mining and content mining. It extracts the data according to user movement and find out the future requests based on these patterns. The textual content of WebPages is extracted using character N-grams, which are integrated with weblog files to derive user navigation profiles. This approach can be used for personalization and better organization of a particular Website. The results obtained provide a better accuracy in terms of Web data extraction parameters. The approach is applied over three types of classifications. The first classification is equal weight methods and finds out better efficiency and accuracy than the method available in literature. Second classification is KNN classification which is evaluated through GM based dissimilarity and performs better in form of classification and prediction. Third classification is based on N-gram size i.e. document frequency and results obtained are again better the techniques used in literature.

Kyung-Joong Kim et.al (2) proposed an approach for Web content mining based user perspective and choice. Hence it provides the content extracted from the combination of Web usage mining and Web content mining. The proposed approach uses fuzzy integral system for this purpose. It uses Structure Adaptive Self-Organizing Map (SASOM) which is a variant of SOM for pattern recognition and visualization. Experimental results obtained by the proposed method are compared with Naive Bayes Classifier method of data mining and are found better in terms of various parameters of Web data extraction.

III SYSTEM MODEL

To provide content available on the Websites according to user choice is a great challenge today. It requires combining the tasks of Web content mining and Web usage mining. The combining process takes a lot of time and provides less accurcy.

III a) N-GRAM MODEL

N-gram model (1) is combined effect of Web usage mining and Web content mining.

 ISSN: 2454-4248
42 – 44

The steps used in this approach are :

Step1 Data cleaning : It is used to extract only the relevent information from the Weblog and rest to be cleaned.

Step2 Web usage mining : It uses clustring approach(Session clustring) for this purpose.

Step3 It combines data obtained in web usage mining with web content mining using **N-grams** method.

Step4 Extract the data based on combination of web usage mining and web content mining.

III b) SYSTEM MODEL (CCMUM)

CCMUM is used to combine the advantages of Web Content Mining and Web Usage Mining. It uses the approaches already defined in previous chapters. CCMUM uses MCMM-LSW model for Content Mining [11] and UMC model for usage mining process. Finally it uses n x 1 to remove the noise [10]. Hence CCMUM acts as a complete web wrapper for efficient web Data Extraction.

The Steps used for CCMUM model are as follows

Input : Website W consisting of n WebPages

Output : Extracted content E from Website W.

Step 1: Develop the webpages into n x 1 format.

Step 2 : Enter the keyword K.

Step 3 : E1=Extracted results by Applying MCMM method according to k.

Step 4 : E2=Extracted results byApplying clustring method for user choice.

Step 5 : E3= Overall results from E1 and E2.

Step 6 : E=Apply XSLT method over E (Noise free results).

CCMUM model is described in Figure 1

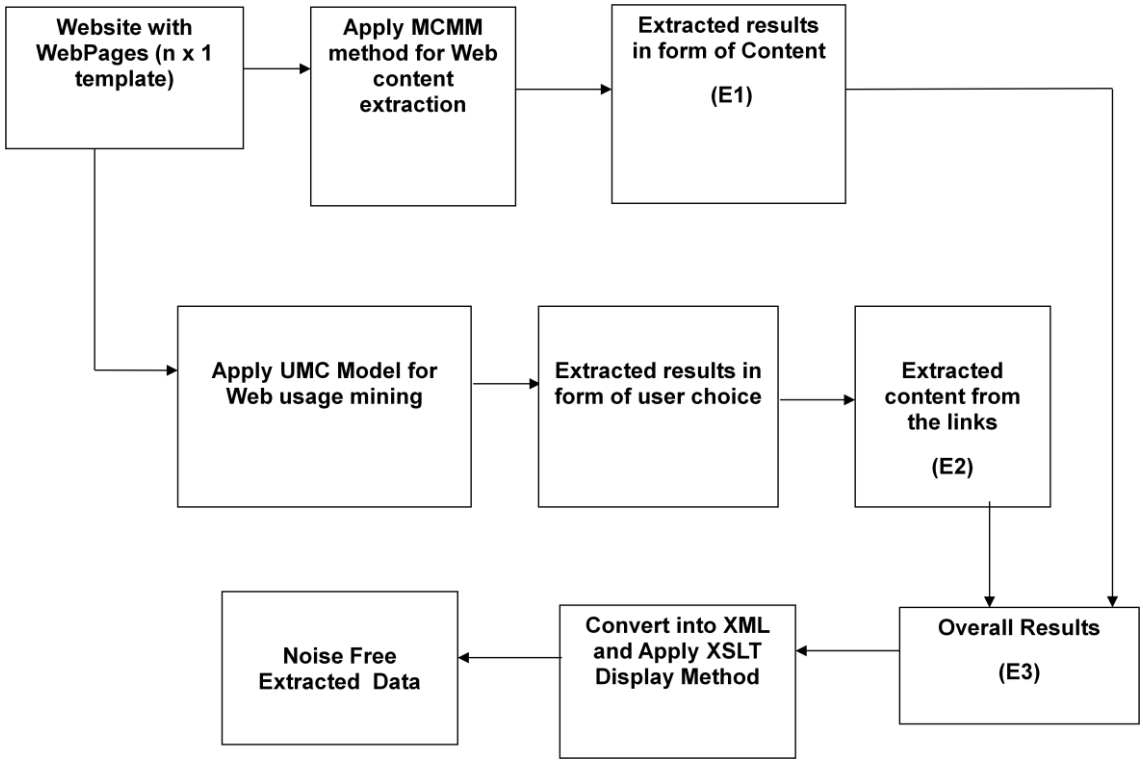

Figure 1: CCMUM Model

CCMUM will use MCMM-LSW model for providing accuracy and both MCMM-LSW and UMC model for less data extraction time.

In CCMUM model step1 of N-gram model (80) i.e. Data cleaning is not required. As cleaning of weblog can be performed very easily. Because in UMC model relevancy rank repor is used to stores the visit count of each webpage or link. Hence if it can be cleaned without considering any details i.e. it can be evaluated at any time using the list developed during the process.

IV EXPERIMENTAL RESULTS AND DISCUSSIONS

For the Experiemental results three Websites named Website1, Website2 and Website3 are developed. The results of CCMUM and N-gram model will be compared on the basis of efficiency parameters for Web data extraction i.e.

 ISSN: 2454-4248
42 – 44

$Precision\ (P) = (LEC - (LEC - LEP + LM))/LEC$
$Recall\ (R) = (LEC - (LEC - LEP + LM))/LEP$
$F - measure\ (F) = 2 * ((P * R)/(P + R)$
$Data\ extraction\ Time = Time\ to\ extract\ data$

Whereby

- LEC refers to extracted content.
- LEP refers to expected content
- LM refers to missing content

CCMUM model is still to be implemented and results will be provided accordingly.

V Conclusion

This research work provides a model CCMUM which combines MCMM-LSW and UMC models and takes advantage of both. It is expected to provide efficient results than the N-gram model in terms of efficiency parameters.

References:

[1]. Liu H, Kešelj V. Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data & Knowledge Engineering. 2007 May 31; 61(2):304-30.

[2]. Kim KJ, Cho SB. Fuzzy integration of structure adaptive SOMs for web content mining. Fuzzy Sets and Systems. 2004 Nov 16; 148(1):43-60.

[3]. Baglioni M, Ferrara U, Romei A, Ruggieri S, Turini F. Preprocessing and mining web log data for web personalization. In AI* IA 2003: Advances in Artificial Intelligence 2003 Sep 23 (pp. 237-249). Springer Berlin Heidelberg.

[4]. Debnath S, Mitra P, Pal N, Giles CL. Automatic identification of informative sections of web pages. Knowledge and Data Engineering, IEEE Transactions on. 2005 Sep; 17(9):1233-46.

[5]. Chen J, Shankar S, Kelly A, Gningue S, Rajaravivarma R. An adaptive bottom up clustering approach for Web news extraction. In Wireless and Optical Communications Conference, 2009. WOCC 2009. 18th Annual 2009 May 1 (pp. 1-5). IEEE.

[6]. Kosala R, Blockeel H. Web mining research: A survey. ACM Sigkdd Explorations Newsletter. 2000 Jun 1; 2(1):1-5.

[7]. Jose J, Lal PS. Extracting Extended Web Logs to Identify the Origin of Visits and Search Keywords. In Intelligent Informatics 2013 (pp. 435-441). Springer Berlin Heidelberg.

[8]. Chau R, Yeh CH. A multilingual text mining approach to web cross-lingual text retrieval. Knowledge-Based Systems. 2004 Aug 31; 17(5):219-227.

[9]. Wang C, Lu J, Zhang G. Mining key information of web pages: A method and its application. Expert Systems with Applications. 2007 Aug 31; 33(2):425-33.

[10]. Neeraj Raheja, V.K.Katiyar " A Noise Reduction Approach based on n x 1 table and XSL display method for efficient web data extraction" , " IJCA International Journal of Computer Applications (0975 – 8887) Volume 64– No.11, pp. 12-17, February 2013.

[11]. Neeraj Raheja, V.K.Katiyar, "MCMM-LSW: Multilevel Content Mining Model for Large Scale Websites", in International Conference on Advances in Information Technology and Mobile Communication – AIM-2015, pp.90-99, August 2015.