XML: URL Data Set Creation for Future Web Mining Research Avenues

Krishna Murthy. A[1]
DoS in Computer Science
University of Mysore
Mysore – 570 006, India

Suresha[2]
DoS in Computer Science
University of Mysore
Mysore – 570 006, India

ABSTRACT
The rapid expansion of the internet has made web a popular place for disseminating and collecting information and also it opens many research topics on varies research fields. Since last few years, several attempts have been made on Web based research particularly based on HTML web pages because of its more availability. So that many Research Data sets have created and few of them are made available on Web. But W3 consortium stated that, HTML does not provide a better description of semantic structure of the web page contents. To overcome this draw back Web developers started to develop Web page(s) on XML, Flash kind of new technologies [1]. It makes a way for new Research methods. This article mainly focuses on Data Set creation on XML Web pages by using Sequential search, Link Extraction and string based classification methods for future research avenues on XML Web pages.

Keywords
URL data set, XML URL's, URL Extraction, URL Classification.

1. INTRODUCTION
Now a day as we all know Research on Web is the emerging field. For example improving the quality of Web by Analyzing Usability Test, Web Information Extraction, Browsing Web on Small Screen Devices [5][6][7][8] like mobile, PDA (Personal Digital Assistance) etc., Tracking Product Opinions by analyzing user reviews such as we can say plenty problems. In general we call it as 'Web Mining'. According to analysis targets [2], Web Mining can be divided into three different types, which are *Web Usage Mining*, *Web Structure Mining* and *Web Content Mining*.

- *Web usage mining* is the process of finding out what users are looking for on the internet (it describes how a page is used, the date and time it was accessed, the IP address of the browser and page references etc.)[9][10].
- *Web Structure Mining* is the process of using graph theory to analyze the node and connection structure of the web site. According to the web structural data, Web Structure Mining can be divided into two kinds; extracting patterns from hyperlinks in the web (a hyperlink is a structural component that connects the web page to a different location) and Mining the document structure (analysis of the tree link structure of page structures to describe HTML or XML tag pages) [11][12].
- *Web Content Mining*, it aims in extracting useful information or knowledge from web page content. Many works have been done and also being carried out on these areas particularly based on HTML, because of more availability of HTML Web pages on Web. So that we get HTML data sets easily [13][14][8].

W3 (World Wide Web) consortium stated that, HTML has a lot of drawbacks such as limited defined tags, not case sensitive, semi-structured and designed for only to display data with limited options. Later to overcome these difficulties few technologies have been introduced such as XML, Flash (with good design options) and so on [1]. Therefore Web developers started to migrate to develop Web pages on these kinds of emerging Web Technologies to provide a better description of semantic structure of the web page contents. Therefore these days we can see more web pages on Web which are developed using XML and Flash technologies [3].

So that there are many research fields have been opened on these new technologies. To pursue research initially we need a Data Set (URL/Web page Collections). Getting .xml extension based URL's is very difficult task because if you do string based search in search engines such as Google, Yahoo etc., (Ex: based on '.xml' query), it will provide you the content based related results. We need semantic structure related results and also existing URL's will not be stored in any servers to fetch directly. Therefore we need a new system/method to perform the task and fulfill the requirements. Here we have introduced the system which extracts all URL's of given web site(s) and classifies XML URL's out of extracted URL's.

2. OUTLINE OF PAPER
In next section architecture of the proposed system has been discussed. Forward by this the entire flow of the proposed model is presented and then algorithm of proposed model has been discussed along with nomenclature. Finally observation of proposed model and conclusion has been discussed.

3. ARCHITECTURE OF THE PROPOSED SYSTEM
The Figure 1 depicts the overall architecture of the proposed system. Here we planned to extract all URL's of given Website and classify the XML URL's of it. In our proposed system, in very first step around 120 million Web Domains (.net, .org, .com, .info, .us, .biz and .sk) are downloaded from Web and dumped them into DB for example www.brainbench.com, www.hollywoodauditions.org and so on .

Then the obtained Web Domains are fetched one by one and given them as input to the URL Extractor. URL Extractor first extracts URL's of main page of given Web Domain (web site) then it works recursively to retrieve entire URL's of Website by getting input of extracted URL's. Finally, by using string based search, XML URL's are classified and stored them into a separate array.

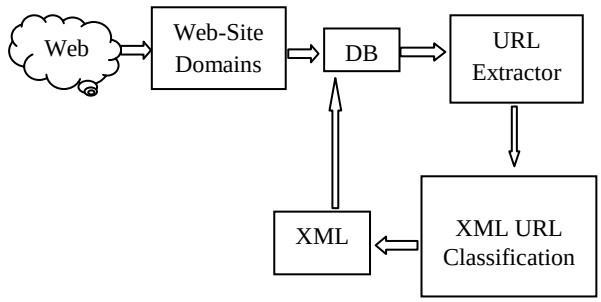

Fig. 1: Architecture of the proposed system

Algorithm recursively continues for the next elements from DB, in case of no element found in DB algorithm will ends. We have extracted more than 2000 XML URL's by using the proposed method.

4. FLOW OF THE PROPOSED MODEL

First we retrieve Web Domain names one by one from DB, input it into Link Extractor. It gives existing URL's of main page of Web Domain; store it into an array called 'M'. In second step read URL's from 'M' meanwhile check conditions whether the end of an array 'M' and domain of URL. If end of an array terminate the URL extraction process or in case of same domain, give URL as input to Link Extractor, after extracting URL's (links of current Web page) store them into array called 'm'.

Now by using sequential search method check for existence in 'M' for each element of 'm' by reading one by one. Upend the URL to 'M' in case of not existence otherwise move to next element until the end of an array 'm'. In general extracted i^{th} URL from 'm' is compared with all URL's of 'M'. If any redundant link is noted, then search will move to 'i+1'th location element or else URL will be upend to the URL list, this process will exit when 'i' reaches to an end of 'm'. The above process continues for all element of the array M. Figure 2 shows the entire flow of the proposed model.

After that by using string based search method XML URL's are identified by searching '.xml' string on each element of an 'M' and found URL's are stored into an another array called 'X'. Once this process ends search will move back to read next element from array 'M' and continues the same process till end of an array 'M'. Therefore we have finally got existing URL's of retrieved Web Domain(s) from DB and classifies the XML URL's. Again from beginning algorithm will works for upcoming Web Domain(s) from DB. Algorithm will ends when retrieving Web Domain reaches to an end of DB.

5. ALGORITHM OF THE PROPOSED SYSTEM

Input : Web Domain D_i
Output: Existing XML URL's
Other Variables: Array – m[], M[], X[]

```
Extract the input Web Domain (D) from DB
M = Extract URL's of first page of Web Domain
for (i=1; i<=End of M; i++)
{
m=Extract URL's of M[i];
for (j=1; j<=End of m; j++)
{
Search m[j] in M
if (not existence of m[j] in M)
Upend m[j] to M;
}
}
j=0;
for (i=1; i<=End of M; i++)
{
if(existence of String(.xml) in M[i])
{
X[j]=M[i];
j++;
}
}
```

Nomenclature:

D_i → Web Domain
DB → Database – collection of Web Domains
M → URL's of input Website (Web Domain)
m → URL's of individual Web page(s)
X → XML URL's

6. OBSERVATIONS

Experimental results ensure that the system analysis, search time and data set creation time gets reduced by using the proposed system. This method is very simple to implement and the proposed method will be more useful for future research avenues.

7. CONCLUSION AND FUTURE WORK

In this article, we found XML and Flash technologies are introduced to overcome the limitations of HTML. We have presented a brief overview and challenges involved in designing a system to extract URL's of entire website and to classify XML URL's for future research avenues. From this study we have observed that few open research problems on XML, Flash kind of new technology Web pages and we can conclude that, this article will helps to improve the Research methods in different kind of ways. Moreover the presented method will be an initial step for all new research avenues on XML Web pages.

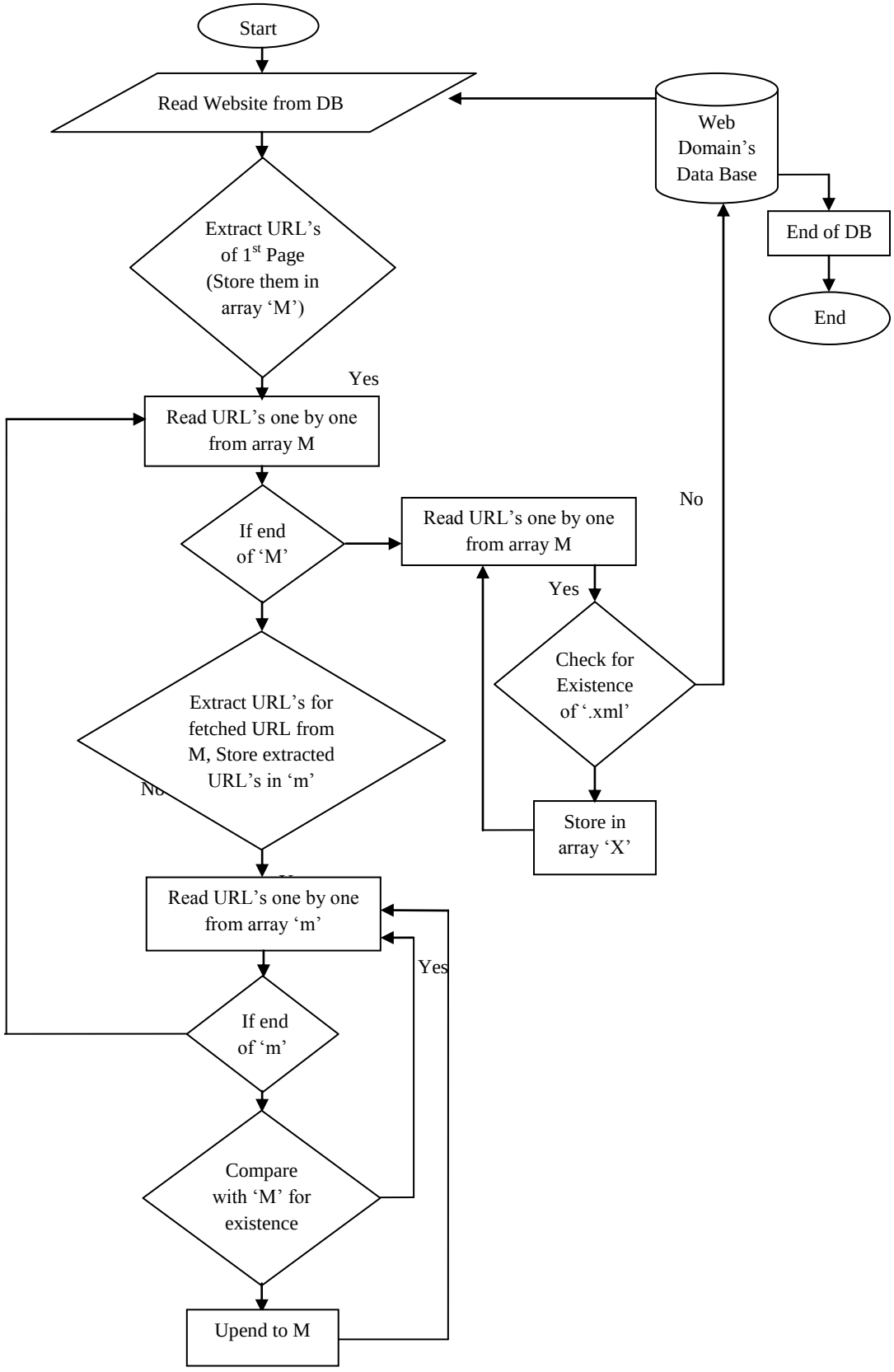

Fig. 2: Flow of the proposed system

8. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

9. REFERENCES

[1] Book: Ed Tittel, 'Complete Coverage of XML', Tata McGraw-Hill Edition.

[2] Book: Magdalini Eirinaki, 'WEB MINING: A ROADMAP'

[3] Lan Yi, Bing Liu, and Xiaoli Li. , 2003, 'Eliminating noisy information in web pages for data mining'. In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 296{305, New York, NY, USA. ACM.

[4] http://www.w3c.org/DOM/

[5] P.F Xiang et al. 2006 'Effective Page Segmentation Combining Pattern Analysis and Visual Separators for Browsing on Small Screens' Web Intelligenc.

[6] Shumeet Baluja 2006, 'Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework'. In WWW '06: Proceedings of the 15th international conference on World Wide Web, pages 33{42, New York, NY, USA. ACM.

[7] Y. Chen, X. Xie, W.-Y. Ma, and H.-J. Zhang, 2005. 'Adapting web pages for small-screen devices' Internet Computing, 9(1):50–56.

[8] Xin Yang, Yuanchun Shi, 2009 'nhanced Gestalt Theory Guided Web Page Segmentation for Moile Browsing' IEEE/WIC/ACM.

[9] Jaideep Srivastava_ y , Robert Cooley, et al, 2000, 'Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data' Volume 1, Issue 2 - page 12, ACM SIGKDD.

[10] Abraham. A, 'Business Intelligence from Web Usage Mining', Journal of Information & Knowledge Management (JIKM), World Scientific Publishing Co., Singapore, Vol. 2, No. 4, pp. 375-390, 003.

[11] Soumen Chakrabarti, 2000, 'Data mining for hypertext: A tutorial survey' Volume 1, Issue 2 - page 1 ACM SIGKDD.

[12] Soumen Chakrabarti, Byron E. Dom et al, 'Mining the Link Structure of the World Wide Web' 1999. _IBM Almaden Research Center, 650 Harry Road, San Jose CA 95120.

[13] C. Kohlsch utter and W. Nejdl. 2008, 'A Densitometric Approach to Web Page Segmentation'. In ACM 17th Conf. on Information and Knowledge Management (CIKM 2008), 2008.

[14] Christian Kohlschutter, Peter Fankhauser, Wolfgang Nejdl, 2010 'Boilerplate Detection using Shallow Text Features', WSDM, New York, USA, ACM.

[15] G. Poonkuzhali, K.Thiagarajan, and K.Sarukesi, 2009 'Signed Approach for Mining Web content Outliers', World Academy of Science, Engineering and Technology 56.

[16] Bar-Yossef, Z. and Rajagopalan, S., 2002 'Template Detection via Data Mining and its Applications'. In Proceedings of the 11th International World Wide Web Conference (WWW2002).

[17] Lin, S.-H. and Ho, J.-M., 2002, 'Discovering Informative Content Blocks from Web Documents'. In Proceedings of ACM SIGKDD'02.