© UNIVERSAL RESEARCH REPORTS | REFEREED | PEER REVIEWED
ISSN : 2348 - 5612 | Volume : 06 , Issue : 01 | January – March 2019

MULTILEVEL CONTENT MINING MODEL FOR LARGE SCALE WEBSITES

Sushil Kumar Sharma
Dr. B.R. Ambedkar Govt. College, Kaithal
E-mail: shilu_online@rediffmail.com

ABSTRACT

As per the current usage of WWW, the data available over the Websites is also growing at a large scale. Hence, efficient Web data extraction has become a great challenge for large scale Websites. The main requirement of a user from such types of Websites is to extract the accurate data in desirable amount of time. This researchwork provides a Web content extraction model for extracting content from large scale Websites. The System Model (MCMM-LSW) produces a link tree of Website and extracts content based on the seed page extracted from different levels of link tree. The results produce higher recall, precision and overall accuracy (F-measure) than the approach used in the literature i.e. 2-level approach.. The effect of applying MCMM-LSW on changing the number of levels of the Websites is shown in the results. Finally the comparison of keyword based extraction and MCMM- LSW is also shown. `

INTRODUCTION

As the number of Websites and data available over these Websites is growing at a rapid rate therefore, Web data extraction systems are required to extract the data according to a particular pattern from these large scale Websites. The data available on the Websites is unstructured or semi-structured in nature thus, the user who is extracting the data faces a lot of problems like extracting accurate, relevant, noise free and user perspective data .

The data available over the Websites is generally distributed in a number of WebPages according to a particular topic or category. So, a particular pattern must be searched from the WebPages to extract useful data such as category of products, common feature information etc.

Web mining which is a type of data mining can be used to extract data from WebPages content according to a particular pattern (Web Content Mining), from WebPages links (Web Structure Mining) or from Weblogs (Web usage Mining) . The focus of this research work is Web content mining. Web content mining is used to extract key information from WebPages . Web content mining works in two modes, data extraction from WebPages and data extraction from database. First mode provides data filtering from WebPages according to user need while in database mode data is extracted on the basis of query executed by the user on a structured database .

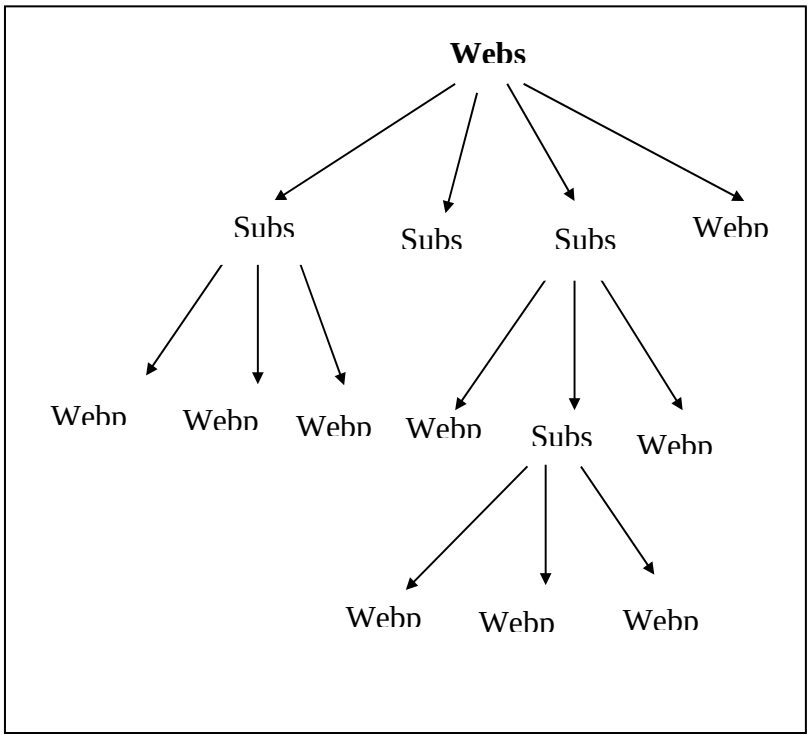
The main requirement of a user who is surfing a popular and large scale Website is to extract accurate data in desired amount of time. This research work presents such a model based on Web content mining to extract accurate results from large scale Websites.

PROBLEM STATEMENT

The data extraction approaches normally uses keyword matching (meta-data) based data extraction techniques. To increase the accuracy of content extracted, content matching based data extraction approaches are required but it takes significantamount of time or applicable over to small scale Websites. Hence, to extract data from large scale Website using content matching so that both accuracy and desirableamount of time can be achieved is a great challenge.

PROPOSED SYSTEM MODEL

MCMM-LSW is used to extract content from both small scale as well as large scale Websites. The model of such a website is shown in Figure 1.1, which consists of multilevel structure. The existing approach uses 2-level approach i.e. the seed



page is index page or default page of the Website (first level) and linked WebPages on this seed page (second level).

Figure 1.1: Structure of a Large Scale Website

A Website comprises of a number of interlinked WebPages and subsites. A subsite is a complete Website in itself which may consist of interlinked WebPages and subsites. It is described as

$$W = \sum_{i=1}^{n} Pi + \sum_{j=0}^{m} Sj \text{ where } Sj \subseteq W \quad \ldots (1.1)$$

(i.e. a subsite represents a complete Website in itself which consist of n WebPagesand m subsites)

Table 1.1 gives the list of notations used along with their description in the largescale Websites.

© UNIVERSAL RESEARCH REPORTS | REFEREED | PEER REVIEWED
ISSN : 2348 - 5612 | Volume : 06 , Issue : 01 | January – March 2019

Table1.1: Notations used for Large Scale Website

Notation	Description
W	Website containing WebPages and subsites
Pi	Set of WebPages
Sj	Subsites in the Website W
N	Number of interlinked WebPages
M	Number of subsites

The working procedure of system model MCMM-LSW is shown in Figure 4.2.

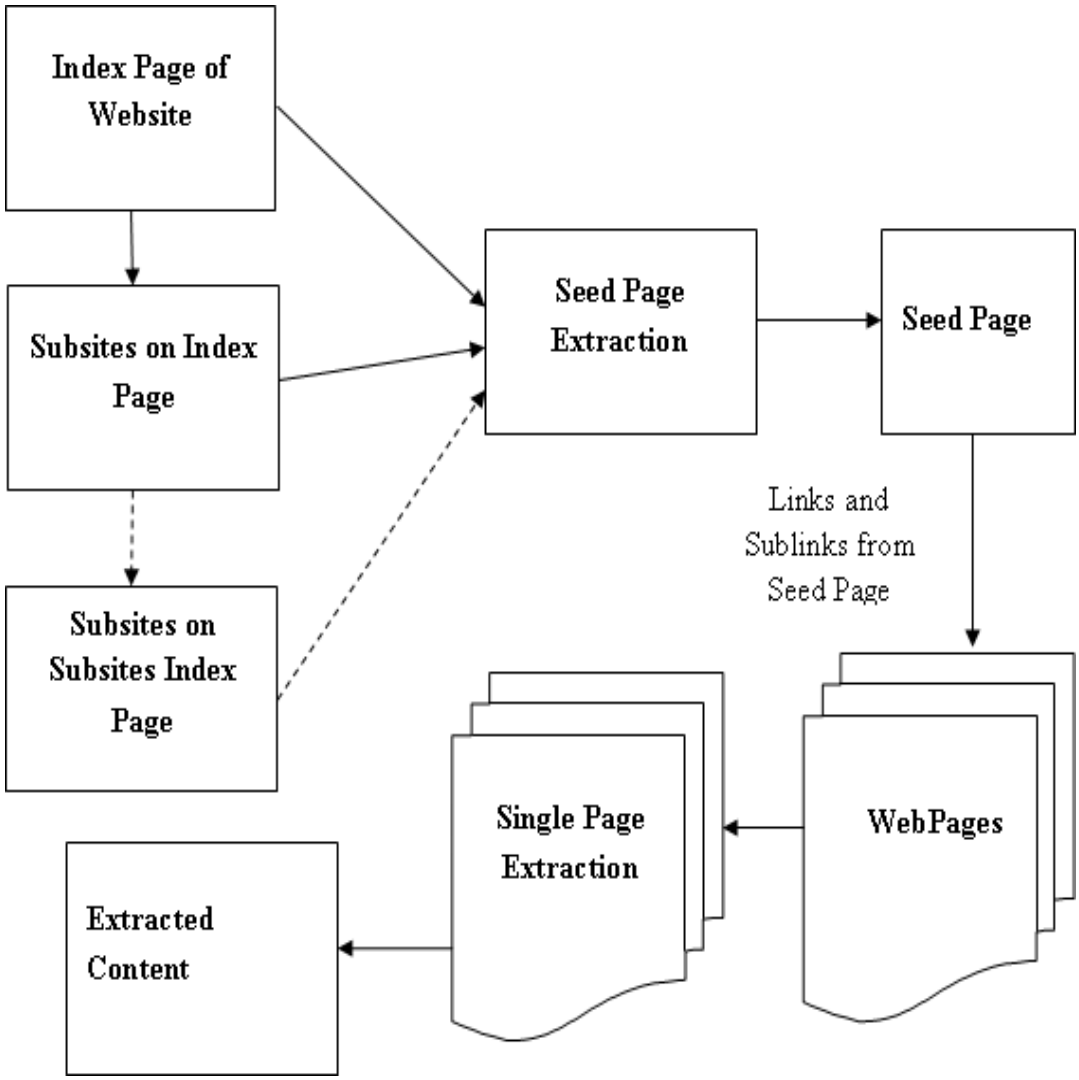


Figure 1.2: MCMM-LSW Model

MCMM-LSW shown in Figure 1.2 works on the basis of following major phases:

Seed Page Extraction: In this phase a link tree is created for the Website as shownin Figure 4.1 from which the content is to be extracted. It is also known asconceptual design of a website. Every website consists of a main webpage or indexpage recognize by the web server as default webpage, subsites and interlinkedWebPages. A subsite is a complete website in itself (created same as a website andconsist of a main or index page as default Webpage, links and subsites). To extractthe content, the seed page is to be first extracted which may be the index page of thewebsite or index page of subsite on website index page or index page of subsite onsubsite index page. If in a subsite the link of any other index page is found then theindex page related to that subsite is considered only. The searched term is matchedwith the content of the index page of website and index page on subsites of indexpage. The seed page is temporarily considered as page with largest match count andthen the process is repeated by considering the temporarily found seed page as mainpage i.e. term is matched with the subsites index pages on temporarily found seedpage. Finally the page with largest count matched is considered as the seed page.

Content Extraction: The content to be extracted is extracted from the seed page found in the seed page extraction process. In this phase the entire link pages and subsite links from the seed page are extracted. The keyword to be matched is matched with all the WebPages content. If the largest content matched is found on seed page then the content of all the linked pages and subsite links is merged and shown as content extracted. If the largest count is found on a particular page then the Webpage content is extracted is treated as content extracted.

Single Page Extraction: In this process the content of the Webpage is extracted by removing the noise from the Webpage like advertisements, links etc.

When MCMM-LSW model described above is applied to any kind of Website i.e. either small scale or large scale Websites, it provides efficient results as the size of the Website will increase.

REFERENCES

[1] Web Data Extraction, Applications and Techniques: A Survey by Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner published at ACM Computing Surveys, Jul 2012.

[2] Yuefeng Li and Ning Zhong: Web Mining Model and Its Applications for Information Gathering, Knowledge-Based Systems 17, pp. 207–217, 2004.

[3] Rekha Jain and Dr. G. N. Purohit,"Page Ranking Algorithms for Web Mining, International

Journal of Computer Applications",ISSN: 0975 – 8887, Volume 13– No.5, pp. 22–25, January 2011.

[4] Claudia Elena DINUCA, "An Application for Data reprocessing and Models Extractions in Web Usage Mining", International Conference on "Risk in Contemporary Economy", Galati, Romania. ISSN 2067-0532, XIIth Edition, 2011.

[5] S. Brin, and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[6] Wenpu Xing and Ali Ghorbani, "Weighted Page Rank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[7] J. Kleinberg, "Authoritative Sources in a Hyper-Linked Environment", Journal of the ACM 46(5), pp. 604-632, 1999.

[8] Cooley R., Mobasher B., Srivastava J. "Web mining: Information and Pattern discovery on the World Wide Web. A survey paper". In Proc. ICTAI-97, 1997.