Automatic Web Data Record Extraction Using AutoRM And Data Path Code Alignment

Hariharan S[1] ,Archana P V[2], Sowmya R[2] ,Subhashree D[2], Varshini M[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore

[2]U.G Student, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore

Abstract - With rapidly growing technology and popularity of the World Wide Web, the difficulty in managing information has become enormous.These information are primarily stored as data records which are retrieved from databases so that they can be viewed in web pages with fixed templates.The proposed system is based on a new approach called AutoRM (Automatic Record Mining) ,which mines data from the webpage automatically and is highly effective. By making more robust assumptions, AutoRM tends to detect data record boundaries more accurately.By constructing a DOM tree and applying various steps as in AutoRM and by utilizing Data Path Code Alignment technique the search procedure is tried to be made efficient and faster than the existing one.

Keywords – AutoRM, Dom tree, Data Path Code Alignment

I. INTRODUCTION

Web Mining is generally employed to extract meaningful information from the web and to find out relevant search results.Generally this kind of data extraction from the World Wide Web is done using certain search engines like Yahoo, Bing, MSN, Google, etc.,Data Mining involves analysing usable information and thereby extracting data from data warehouses involving different algorithms, patterns and tools.This process of data extraction can be used for data analysis, predicting user behaviour and future trends.

The data are generally stored in a Data-Wabehouse, which includes the objective of construction ,which is to collect specific data to analyze the behavior of navigation

The data used can be classified in to four types :[2]

- Content Data: Data contained in the Web pages (texts, images, graphics…)
- Data relating to the structure: structure of the page, structure inter-page
- Data relating to the Use: data providing information on the use such as IP addresses, the date and the time of queries.
- Data relating to the profile of the user

Types of Web Mining[1]

Web mining can be broadly divided into three different types of techniques.

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

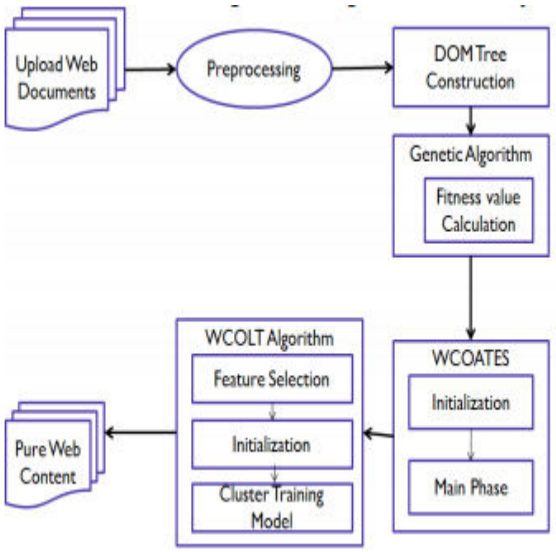

Fig 1.1 System Architecture for Clustering records [3]

Approaches in Web content mining

- Agent-based approach
- Data-based approach

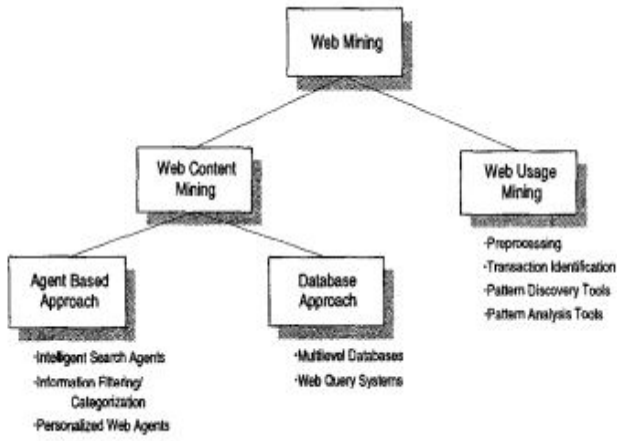

Fig 1.2 Web mining Classification

A. Web Content Mining

The Web content mining is the mechanism for extracting knowledge from the actual content of documents and web pages such as structured recordings, images, texts, videos, etc [7].

B. Web Structure Mining

Web Structure Mining is an analysis of the structure of Web i.e. the architecture and the links that exist between the different sites. The analysis of the paths travelled allows, for example, determining how many pages to consult the internet users on the average and thus adapt the site tree for that the pages of the most sought after are in the first pages of the site. Similarly, research associations between the pages consulted allows improving the ergonomics of the site by creation of new links. We cited two well-known structures mining algorithms, PageRank and HITS [7].

C. Web usage mining

The purpose of this methodology is to analyze user behavior and to extract interesting usage patterns from the interaction with the website. There are three steps in web usage mining that helps to navigate the web in an effective way. These steps include the preprocessing, pattern discovery and pattern analysis.

Techniques used in Web Content Mining

- Unstructured Data Mining
- Structured Data Mining
- Semi – Structure Data Mining
- Multimedia Data Mining

Tools:

- Screen Scraper
- Mozenda
- Automation Anywhere7
- Web Content Extractor
- Web Info Extractor
- Rapid Miner

Algorithms:

- Naive Bayes
- Decision Tree
- Support Vector Machine
- Neural Network

AutoRM

AutoRM mines data records by employing the following steps.[1]

1. It constructs the DOM tree of the given Web page.
2. Mining each set of adjacent similar C-Records (Candidate data Records) fromthe constructed DOM tree: Here, similar C-Records are actual similar data records or objects that contain smaller actual data records. In many Web pages, similar data records are distributed among multiple data regions embedded in bigger adjacent similar objects.
3. Mining actual data records from each set of similar C-Records.

Data Path Code Alignment

After constructing the DOM[3] tree, Data Path code Matching (DPM) is introduced to identify sub-trees with similar structures and then classify these sub-trees into clusters. These clusters (formed either by combining or deleting the identified structures) represent different sets of data records. It then filters out unimportant sets of data records and manipulates extracted objects from nested data regions. After doing the above steps, a method called Data Path Code Alignment is being employed for extracting corresponding data items from those extracted objects.

Architectural Design

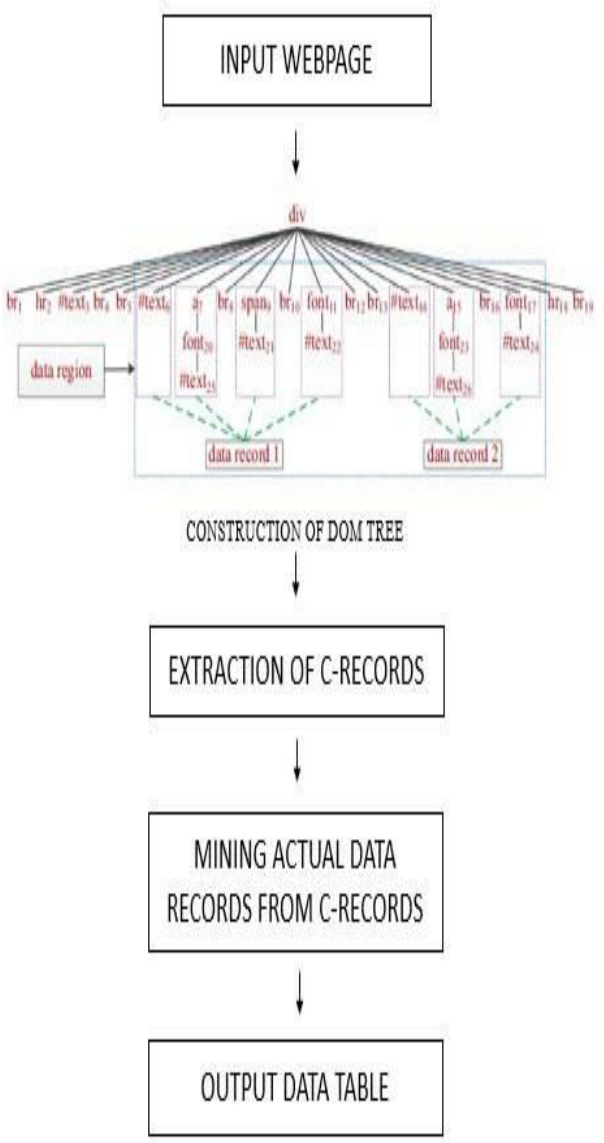

Fig 1.3 Proposed System Design

Work Done

In this paper, we have demonstrated a web crawler which initially extracts all the relevant links for the product queried by the user. The figure below gives an overview of how the scraping of relevant urls take place.From the extracted set of individual urls,the required product details are extracted using suitable web

crawling tools and displayed in the webpage for the users.

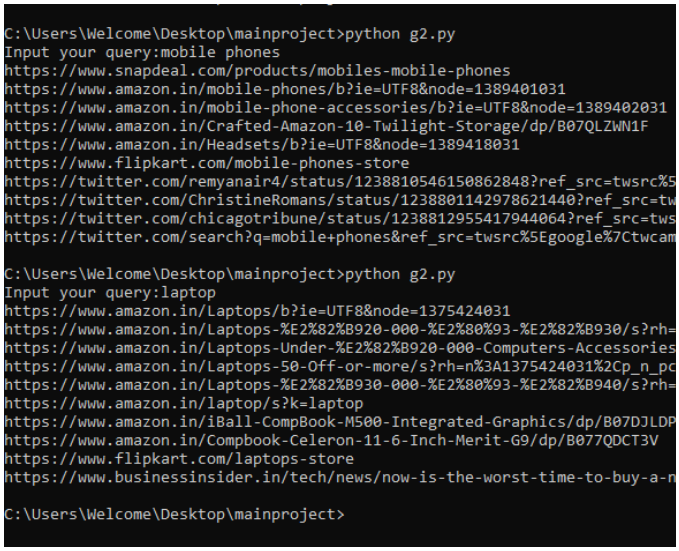

Fig 1.2 URL Extraction

Conclusion

In this paper a novel idea for providing the users with efficient search results in an aligned manner has been proposed.With the details extracted in this procedure and by further making use of additional alignment algorithms and certain efficient web crawling tool, the data retrieval and display can be made effective.

Future Works

By applying Data Path Code Alignment technique,the extracted data from the webpage will be aligned based on a specific criteria which effectively reduces the complexity involved in searching, thereby making the search results more efficient.

References

[1] Automatic Data Extraction of Websites Using Data Path Matching and Alignment by Yu-Chun Chu, Chiun-Chieh Hsu* , Chen-Jhe Lee, Yu-Ting Tsai ,2015, IEEE

[2] Web mining techniques and applications: Literature review and a proposal approach to improve performance of employment for young graduate in Morocco,2018, IEEE

[3] Analytics of Noisy Data in Web Documents Using a Dom Tree, IJARCSSE,2015

[4] AutoRM: An effective approach for automatic Web data record mining Shengsheng Shi, Chengfei Liu , Yi Shen, Chunfeng Yuan , Yihua Huang ,2015

[5] ClusteringWeb Pages Based on Structure and Style Similarity, Thamme Gowda, Chris Mattmann,2016,IEEE

[6] Web Mining: Information and Pattern Discovery on the World Wide Web , R. Cooley, B. Mobasher, and J. Srivastava

[7] M. Balabanovic, Yoav Shoham, and Y. Yun. An adaptive agent for automated web browsing. Journal of Visual Communication and Image Representation, 6(4), 1995.