Contents lists available at www.innovativejournal.in

Asian Journal of Computer Science and Information Technology

Journal homepage: http://www.innovativejournal.in/index.php/ajcsit

ISSN 2249 - 5126

AN EFFICIENT METHOD FOR WEB DATA EXTRACTION USING PARTIAL TREE ALIGNMENT ALGORITHM

M. Karthikeyan*, P. Aruna

Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India.

ARTICLE INFO

Corresponding Author:
M. Karthikeyan
Department of Computer Science and Engineering, Annamalai University, Annamalai nagar, Chidambaram, Tamil Nadu, India.
mkshkarthik@yahoo.co.in

Keywords: Web Mining, Partial Tree Alignment Algorithm, Meta Tag, URL-Oriented data extraction model (UODE), Tag extraction, Content extraction.

ABSTRACT

With the explosion of the World Wide Web, a wealth of data on many different subjects has become available online. Usually, users retrieve Web data by browsing and keyword searching. But, these traditional methods have their limitations and disadvantages. Search engine helps to retrieve the relevant web sites based on the keyword specified by the user. It performs various operations such as crawling, indexing etc. It displays thousands of links as a result of the web search, but there are many road blocks that can make this process difficult or even impossible. So, the proposed system mainly aims to eradicate the disadvantages of search engines by exploring the contents of a web page to a maximum extent. It finds the exact keywords that match a page. When the search engine searches for web pages related to exact keyword, it can return only a few pages which are highly focused, specific and relevant to the topic. By this, the end-user gets the required information related to the search. Experiment shows that new approach is feasible and effective.

INTRODUCTION

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. There are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web Content Mining is the process of extracting knowledge from the content of documents or their descriptions. Web Structure Mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, Web Usage Mining, also known as Web Log Mining, is the process of extracting interesting patterns in Web access logs.

Web Crawlers are programs that exploit the graph structure of the web move from page to page. The key motivation of the web crawlers has been to retrieve web pages and add them or their representations to a local repository. Such a repository may then serve particular application needs such as those of web search engine.

The problem of extracting data from a Web page that contains several structured data records. The Objective is to segment these data records, extract data items or fields from them and put the data in a database table. There are two algorithms for the data extraction i.e. Top-down, bottom-up algorithm. On the basis of these two algorithms, there is a development of Hybrid algorithm called Bi-Direction Data Extraction. It can be able to extract and discriminate the relevance of different repetitive information contents with respect to the user's visual perception of the web page.

Another method to extract useful information from web pages is, first, extract URLs from web pages and then use these extracted URLs to retrieve next pages via the HTTP request. If all pages are accessed via URLs, such a data extraction model is called the URL-oriented data extraction model. The keywords and Meta tag is intended to provide search engines, when navigating a web page, with a list of words which assist in retrieving the web page when someone searches the web using one or more of those keywords. Keywords and Meta Tag Generator assists search engine in optimization by automating the creation of a keywords and Meta tag based on the words most often used in the page and the way they are used. But the process of selecting keywords automatically requires more fine-tuning than simply identifying the most-commonly occurring words, which may be irrelevant words such as "that", "with" and "other" etc (stop words).

So, the proposed method concentrates on the following points to improve the efficiency of Web data extraction.

1. To improve the efficiency of the Search Engine.
2. To break up various tags in the web page and understand the contents of the web page.
3. To retrieve and extract the hyperlink from the web page.
4. To retrieve and extract the data from the web pages.
5. To retrieve the keywords from the given web pages.

The structure of this paper is as follows: Section 2 discusses some related research work regarding Web data extraction methods. Section 3 describes how data is extracted from Web pages using Partial tree alignment

algorithm. The experimental results are given in section 4. Finally some conclusion and discussion is given in section 5.

RELATED WORKS

The process of information extraction from Web is both interesting and challenging, which could be helpful in Web Searching, Information Retrieval and Web Mining. Web pages on many Web sites are produced dynamically as structural records. Tak-Lam Wong and Wai Lam [1] proposed an unsupervised learning framework which can jointly extract information and conduct feature mining from a set of Web pages across different sites. Important characteristic of this model is that it allows tight interactions between the tasks of information extraction and feature mining. Xiangwen Ji, Jianping Zeng, Shiyong Zhang, and Chengrong Wu [2] devised a new method based on Tag tree template. Web pages from different Web sites are parsed into Tag trees, and then templates of each site are generated from the trees by using a cost-based tree similarity measurement. The exclusive content in each page is then extracted by using the templates to parse the page.

Yewei Xue, Yunhua Hu, Guomao Xin, Ruihua Song, Shuming Shi, Yunbo Cao, Chin-Yew Lin, and Hang Li [3] is concerned with automatic extraction of titles from the bodies of HTML documents (Web pages). Titles of HTML documents should be correctly defined in the title fields by the authors; however, in reality they are often bogus. Manuel Álvarez, Alberto Pan, Juan Raposo, Fernando Bellas and Fidel Cacheda [4] proposed a set of novel techniques to avoid multiple inputs, while several previous works have addressed the same problem; most of them require multiple input pages. Gerd Stumme, Andreas Hotho, Bettina Berendt [5] conducted a survey which analyzes the convergence of trends from Semantic Web and Web mining. More and more researchers are working on improving the results of Web mining by exploiting semantic structures in the Web, and they make use of Web mining techniques for building the Semantic Web. These techniques can be used for mining the Semantic Web itself.

Gilles Nachouki, Mohamed Quafafou [6] explained the process for mashing heterogeneous data sources based on the Multi-data source Fusion Approach (MFA). The aim of MFA is to facilitate the fusion of heterogeneous data sources in dynamic contexts such as the Web. Viktor de Boer, Maarten van Someren and Bob J. Wielinga [7] developed an approach, that given existing ontologies, extracts instances of ontology relations, a specific subtask of ontology population. They used generic, domain-independent techniques to extract candidate relation instances from the Web and exploit the redundancy of information on the Web to compensate for loss of precision caused by the use of these generic methods. Jer Lang Hong, Eu-Gene Siew and Simon Egerton [8] developed a non-visual automatic wrapper which questions the need for complex visual based wrappers in data extraction. The novel techniques for their wrapper are filtering rules to detect and filter out irrelevant data records, a tree matching algorithm using frequency measures to increase the speed of data extraction, an algorithm to calculate the number and size of the components of data records to detect the correct data region, a data alignment algorithm which is able to align iterative and disjunctive (optional) data items and a data merging and partitioning method to solve the imperfect segmentation problem.

Lirong Wan, Xinjun Wang and Congcong Chen [9] proposed a spiral-decoding method to synchronize the extractions by spiral decoding and the algorithm to realize it. Hua Wang, Yang Zhang [10] presented a Web data extraction method based on simple tree matching by analyzing the structure and content of Web documents. Jellouli I., Mohajir M.E [11] devised an approach that does not make any prior assumption on the design and the format of web pages, it is totally independent and it is able to achieve semantic extraction from a single web page with a single instance. Hao Han, Tokuda T [12] proposed a method for Web information extraction to generate the virtual Web service functions from Web applications at client side. They shows that the general Web applications can be also integrated easily.

WEB DATA EXTRACTION USING PARTIAL TREE ALIGNMENT ALGORITHM

Proposed system consists of three modules. They are

- Tag Extraction
- Content Extraction
- Display content.

1 Tag Extraction

Tag Extraction is the first module in the proposed system. It deals with extracting the tags automatically from the web pages. It requires identifying the HTML source of the web page then separating the tags and the content. Finally the tags are extracted separately. Each Tag is identified separately and the weightage is assigned to the tags by using the Partial tree Alignment algorithm. The objective of this algorithm is to segment the data records, extract data items/fields from them and put the data in a database table. It consists of identifying individual data records in a page and aligning and extracting data items from the identified data records. Partial alignment aligns only those data fields in a pair of data records that can be aligned (or matched) with certainty, and make no commitment on the rest of the data fields. This approach enables very accurate alignment of multiple data records.

Each tag in the source is extracted separately. The tags include Meta tag, Title tag, Attributes, comments and other tags. The weightage is assigned to the words based on the tags. Suppose if the word is present in the META tag means it was given the weightage of 10. The TITLE tag and the LINK were assigned the weightage of 9. Figure 1 shows the block diagram of tag extraction module. Figure 2 shows the steps involved in partial tree alignment algorithm. Tags and its corresponding weightage are given in table 1. Page scraping is done on the source code. It separates the tags and send them to the corresponding packages in which the operations are performed separately. The contents are separated from the tags using the pre defined class called page scraper.

PageScraper sc = new PageScraper(html)

Tags	Weightage
Meta	10
Title & Links	9
H1	7
H2	6
H3,B & I	5
H4	4
H5, Table	3
Others	1

Table 1. Tags and Weightage

2. Content Extraction

Content extraction is the second module. It deals with extracting the contents from the Web pages. Content extraction is the main task. It is done along with the first module, Tag extraction. The Web page contains the information i.e. the data which is to be extracted, and these data are called as interesting data. The interesting information may also be called as the knowledge content of the Webpage. The content gives the details about the Web page. The content is built with various key words. Weightage is assigned to the tags. Each word is separated and the frequency of each word is calculated separately. Then the value of each word is calculated by using the formula Word Value= (frequency)*(weightage). The value calculation is based on the predefined weightage assigned to the tags. Then the priority is assigned to the key-words based on the highest value. This process is called as parsing. The noisy data present in the content such as and, where, when, though etc (stop words) are eliminated. Figure 3 shows the block diagram for content extraction module.

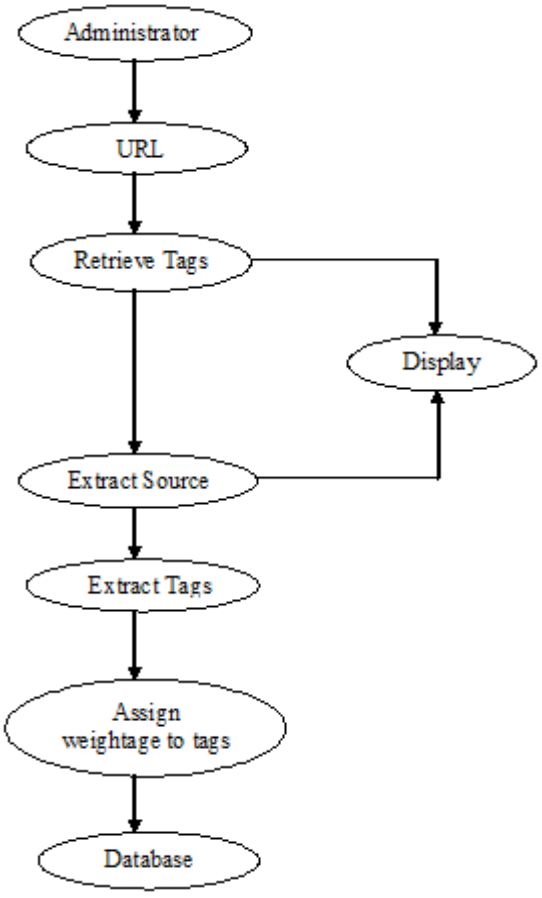

Figure 1. Tag Extraction Module

Let partial tree alignment (S)

1. Sort trees in S in descending order according to the number of Data items that are not aligned;
2. Ts = the first tree (which is the largest) and delete it from S;
3. Flag = false; R = 0; I= false;
4. While (S ≠0)
5. Ti = select and delete next tree from S;
6. Simple_tree_matching(ts, ti);
7. L = align trees(ts, ti); // based on the result from line 6
8. If ti is not completely aligned with ts then
9. I = insert into seed(ts, Ti);
10. if not all unaligned items in Ti are inserted into Ts then
11. Insert Ti into R:
12. endif;
13. endif;
14. if (L has new alignment) or (I is true) then
15. flag = true
16. endif;
17. if S = 0 and flag = true then
18. S = R; R = 0;
19. flag = false; I = false
20. endif;
21. endwhile;
22. Output data fields from each Ti to the data table based on the alignment results.

Figure 2. Steps of Partial Tree Alignment Algorithm

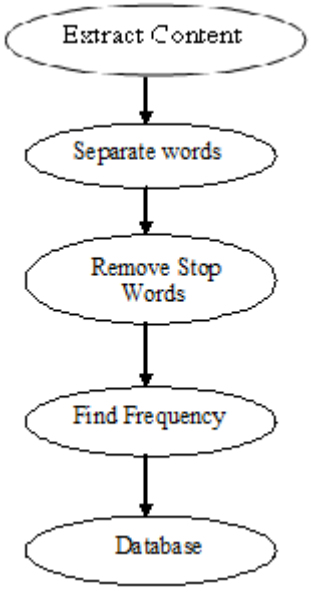

Figure 3. Block diagram of Content Extraction Module

3. Display Content

This module deals with the user interface. It displays the page which is tested (given as input). It also displays the source code, content, links and the ranked key-words. Figure 4 shows the block diagram of display content module.

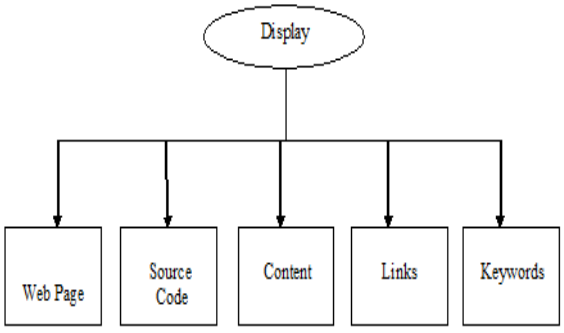

Figure 4. Block diagram of Display Content Module

EXPERIMENTAL RESULTS

The proposed method was implemented in visual basic 6.0 as front-end and oracle as the back-end. The user has to give the URL of the Web page to be tested as input. For experiments, totally ten Websites are considered. All the information are retrieved from the Web pages. Tags, Words, keywords, Hyperlinks are extracted from all the web pages. The URL can be given from the local system or browse through the web. The Web page can be tested both at online and offline. Figure 5 shows the keyword extraction phase and figure 6 shows the hyperlink extraction phase.

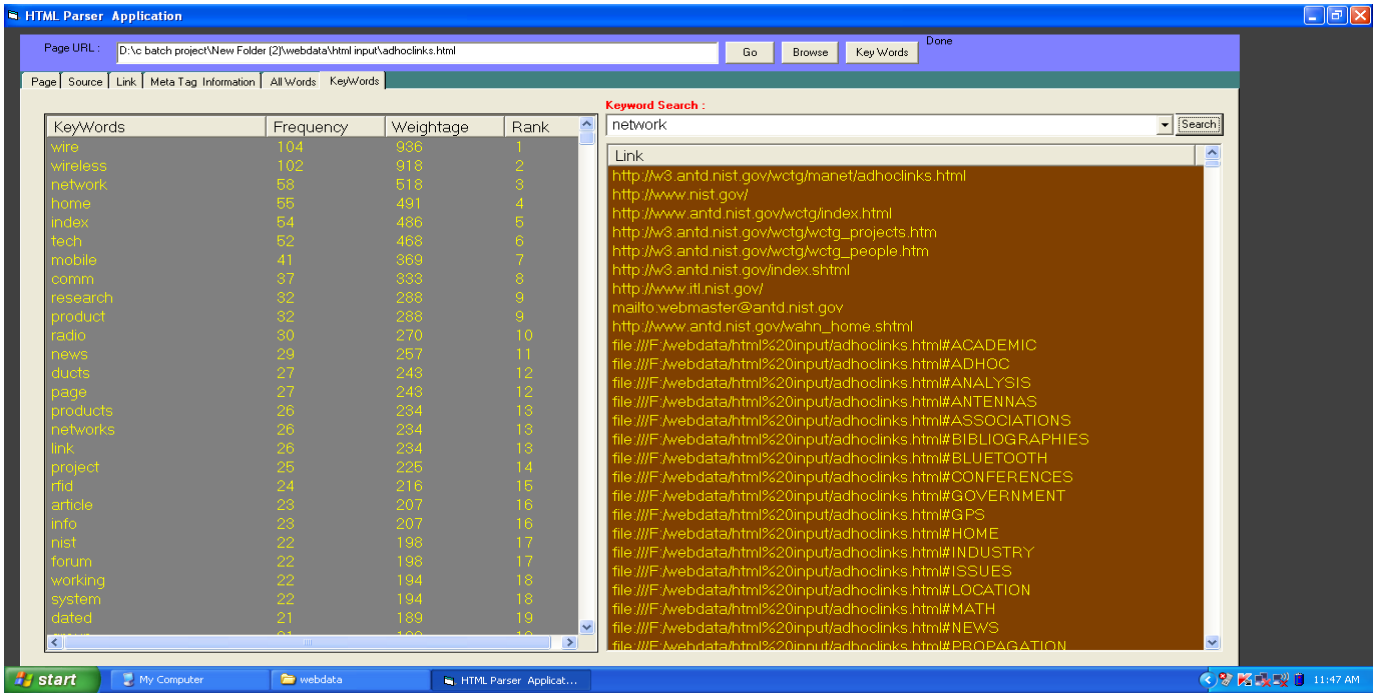

Figure 5. Keyword Extraction Phase.

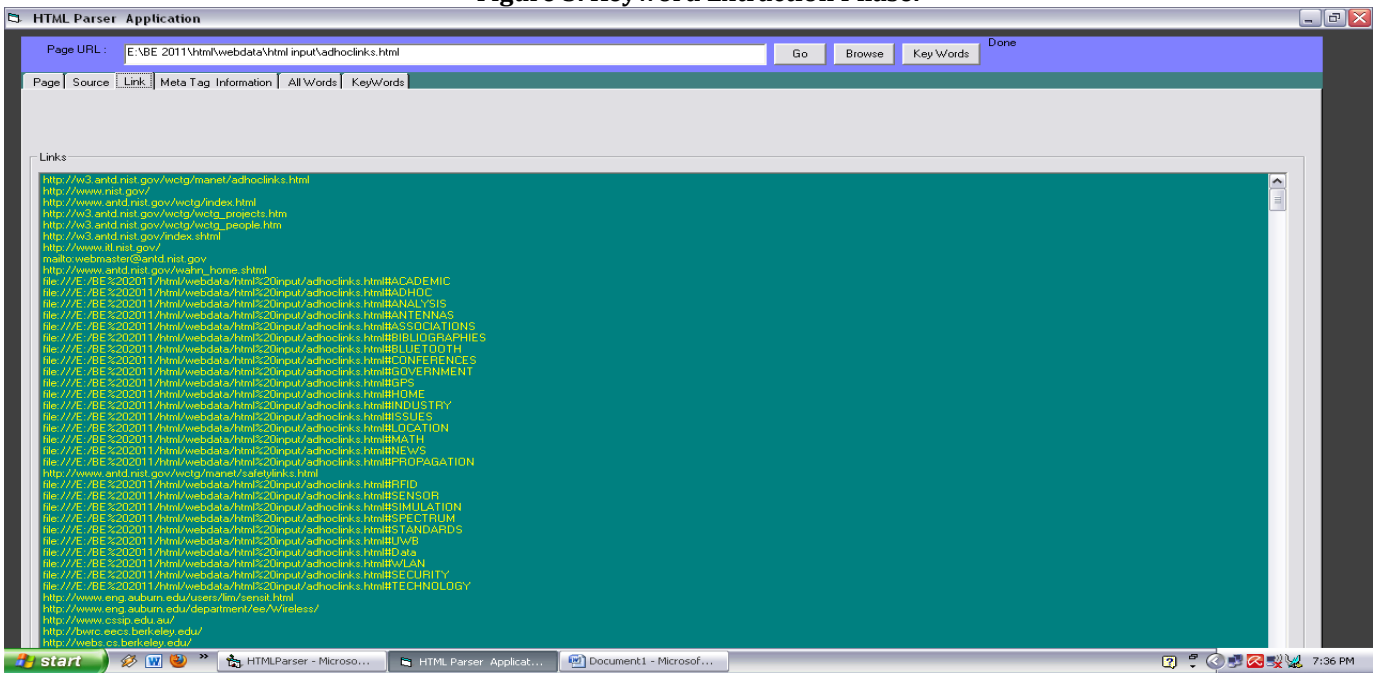

Figure 6. Hyperlink Extraction Phase.

CONCLUSION

The contents of the Web Page were extracted which includes the source code, hyperlinks, Meta tags and keywords. The weightage were assigned to words based on the tags and the ranking was also given by calculating the frequency. The links were displayed based on the keywords. The main goal of the proposed system is based on extracted keyword frequency; hyperlinks may be created in future to provide a convenient way for users to retrieve related documents. In future, this method can be combined with search engine for optimizing it .It can be implemented in other languages. The Extraction of images may be included. Depth of searching can be extended for each linked page.

REFERENCES

1. Tak-Lam Wong, Wai Lam, “An unsupervised method for joint information extraction and feature mining across different web sites”, Data & Knowledge Engineering, Volume 68, Issue 1, January 2009, Pages 107-125.
2. Xiangwen Ji, Jianping Zeng, Shiyong Zhang, Chengrong Wu, “Tag tree template for Web information and schema extraction” , Expert systems with Applications, Volume 37, Issue 12, December 2010, Pages 8492-8498.
3. Yewei Xue, Yunhua Hu, Guomao Xin, Ruihua Song, Shuming Shi, Yunbo Cao, Chin-Yew Lin, and Hang Li, “Web page title extraction and its application”, Information Processing & Management, Volume 43, Issue 5, September 2007, Pages 1332-1347.
4. Manuel Álvarez, Alberto Pan, Juan Raposo, Fernando Bellas and Fidel Cacheda, “Extracting lists of data records from semi-structured Web pages”, Data & Knowledge Engineering, Volume 64, Issue 2, February 2008, Pages 491-509.
5. Gerd Stumme, Andreas Hotho, Bettina Berendt, “Semantic Web Mining: State of the art and future directions”, Web Semantics: Science, Services and Agents on the World Wide Web, Volume 4, Issue 2, June 2006, Pages 124-143.

6. Gilles Nachouki, Mohamed Quafafou, " MashUp web data sources and services based onsemanticc queries", Information Systems, Volume 36, Issue 2, April 2011, Pages 151-173.
7. Viktor de Boer, Maarten van Someren, Bob J. Wielinga, "A redundancy-based method for the extraction of relation instances from the Web", International Journal of Juman-Computer studies, Volume 65, Issue 9, September 2007, Pages 816-831.
8. Jer Lang Hong, Eu-Gene Siew, Simon Egerton, "Information extraction for search engines using fast heuristic techniques", Data & Knowledge Enginering, Volume 69, Issue 2, February 2010, Pages 169-196.
9. Lirong Wan, Xinjun Wang, Congcong Chen, "A Sprial-Decoding Method for Web Data Extraction", IEEE Conference Proceedings on First International conference on Education Technology and Computer Science, 2009, Volume 1, Pages 1026-1029.
10. Hua Wang, Yang Zhang, "Web Data Extraction Based on Simple Tree Matching", IEEE Conference Proceedings on International Conference on Information Engineering, 2010, Volume 2, Pags 15-18.
11. Jellouli I., Mohajir M.E, "An ontology-based approach for Web Information extraction", IEEE conference Proceedings on Information Science and Technology, 2011, pages 5.
12. Hao Han, Tokuda T, "A method for Integration of Web Applicatons Based on Information Extraction, IEEE Conference Proceedings on Eighth International Conference on Web Engineering, 2008, Pages 189-195.