A Study on Web Data Extraction Techniques

D.Thennarasi, S.Krishna Anand
1PG Student, School of Computing (CSE), SASTRA University, 613402,Thanjavur, Tamilnadu, India.
2Senior Assistant Prof, School of Computing (CSE), SASTRA University, 613402,Thanjavur, Tamilnadu, India.

ABSTRACT

A large amount of information is available in the web today. Information extraction is defined as the automatic extraction of structured information from unstructured documents. Even though the web pages are more robust and flexible, the information extraction system transforms the code into user friendly structures. The automated extraction of data from the web pages is still under research. The following study reveals the various techniques involved in the arrangement and alignment of particular information from the websites by eliminating the irrelevant information.

Key words: Data extraction; Parsing; Clustering; Crawler

Introduction

Information extraction is the process of collecting information from the available web database. Web data extraction can be done by various techniques such as fully automated approach or semi-automated. The fully automated approach does not require human intervention. However, Semi automated approach requires some training. The main focus of the extraction is to remove the irrelevant and the redundant data available. The irrelevant data may consist of navigation panel, popup windows and advertisement. Web programming languages are used to extract data from the web database. This provides a qualitative analysis of web data extraction.

II. Literature Review:

Hung-yukao et al. in the year 2004 discovered Lamis and infodiscoverer algorithm. Lamis method is mainly used to collect data from the deep web pages. Info discoverer is applied on the distilled data to obtain a correct precision value. Lamis and infodiscoverer are used to mine information from the news website.

James caverlee et al during the year 2005 proposed thor framework for clustering of pages which are similar. A subtree filtering mechanism is suggested to exploit the particular data from the web database.

Nikolas k padapakis et al.in 2005 disclosed that STAIVES can be operated with less human intervention for extracting the particular data. It uses signal wise treatment of data. Data is arranged in a hierarchical fashion according to a certain priority.

Yanhongzhai et al. in the year 2006 developed Data Extraction based on Partial Tree Agreement (DEPTA) method which is used to perform the data extraction automatically. The desired records in the page are identified. Alignment operations are being applied to data.

Weliu et al. in 2010 explored the revision measurement used by vide to identify the amount of human effort required to perform extraction. It can process webpages with one data region even though multiple web Pages are available.

Luis tari et al. in the year 2012 mentioned that incremental information extraction is a pipeline based approach to extract data from web data source. The extraction process has been carried out for medical databases.

Weifengsu et al. during the year 2012 inferred that Combining Tag and Value Similarity (CTVS) is used to perform three important functions such as data region identification, segmentation of the record and alignment of records.

III. Web Data Extraction Techniques:

3.1. Design of Lamis algorithm:
Over a period of time, it has been observed that news websites contain plenty of relevant information as well as irrelevant information. Hung-Yu Kao et al. designed an algorithm named lamis (Hung-Yu Kao, et al., 2004). This algorithm has the potential to pick out only the relevant information. Besides discovery of relevant information, the algorithm plays an additional role in eliminating the redundancy among the text contents present in the hyperlink induced topic search (HITS) algorithm. HITS are mainly used to calculate the amount of hubs in the given web pages. Lamis uses Eigen vector calculation to obtain the specified result format. The common features are identified using intersection of the desired and the discovered features. This method involves a considerable amount of automation. Two type of redundancies appear in the webpage namely intrasite and intersite redundancy. The intrasite redundancy includes the navigation, control panels and advertisement. Hypertext induced search algorithm and the google page rank algorithms analyse the structure of the web page. The system architecture mainly contains a web crawler and the information extraction module popular for identifying the web contents.

3.2. Introduction to THOR framework:

James Caverlee et al. (2005) proposed a framework named THOR. This framework deals with novel page clustering and filtering mechanism. This algorithm deals with a collection of deep web pages. Initially the web pages with relevant information are collected. This is followed by segmentation of the information. The next step deals with identification of data using page ranking algorithm. The fourth phase deals with separation and localisation of the data utilised. Term frequency inverse document frequency (TFIDF) method calculates similarity, identifies higher page rank list and provides better precision and recall factor. The tag tree based signature uses k-means clustering for identifying the list with higher priority. The user provided query is searched in the corresponding web data bases and the cleaned information is obtained. Tag extraction related methods are carried out on the top ranked clusters. The resultant data enters the single page filtering and cross page filtering. The corresponding query answer page let is selected and remaining information is passed to the deep web service.

3.3. STAIVES Approach:

Nikolaos K. Papadakis et al. (2005). suggested that Stavies is based on the robotic approach for extraction of online data. A semi automated approach has been proposed. This approach deals with analysis and understanding of data. The raw tag calculation has been avoided. The web pages are provided as input and the structural tokens are obtained as output. Extraction of data involves two phases. Preparation phase mainly deals with the hypertext mark up language generation. This is followed by transformation into the tree structure. From the tree structure, the terminal nodes can be determined. In the segmentation phase, the boundary selection and node comparison takes place. The particular token is then provided to the end user for evaluation. The Stavies is a reliable algorithm when compared with mining data record algorithm. Mining data record algorithm is applied mainly to specify web pages whereas Stavies can be applied to deep web pages.

3.4. Roadrunner technique for data extraction:

Yanhong Zhai et al. (2006) suggested that the document object model technique (DOM) is used to identify visual information and tree matching. The nested tag structures are used to build the document object model tree. DEPTA and road runner concepts have been used. DEPTA method could also be used to perform data extraction. However, the problem with this method is that adjacent nodes will not be considered. To overcome this drawback, Road runner technique is used. This technique is based on parsing the data to identify if there is any mismatch in the string. Recursion concept is used to find the matching string. Wrapper technique mainly focuses on the positive result. The system architecture of DEPTA contains four steps namely dom tree builder, data region identifier, data record identifier and data item extractor. The dom tree uses application program interface to retrieve the html documents. Data region identifier is used to identify the similar data records that are available. Data record identifier deals with the measurement of boundaries. Data item extractor is used to align the data in excel format.

3.5. ViDE:

Wei Liu et al. (2010). suggested that Vide technique tries to reduce human intervention while searching for data in the web. There are many kinds of information extraction such as visual based, content based, layout based and appearance based. The noise block which is present inside data is filtered and the remaining blocks are arranged according to the visual similarity. Vide concept mainly includes two components namely viDRE and VIdie. Vide has two drawbacks associated with it. It has been observed that extraction of data is
excruciatingly slow. Besides, it can handle only one data region at a time. Vide is based on the concept of application programming interface. The process of extraction consists of three steps namely filtering of noise blocks, clustering of data based on the similarity data and record recovery. Record recovery could be performed by rearranging the blocks. A new factor revision is applied to calculate the data extraction tools.

3.6. Parse tree generation:

Luis Tari et al. (2012), specifies that parse tree query languages (PTQL) have been used to extract the data. The extraction goals are specified by the user itself. In case any new extraction goal is identified, the parsing is done on the entire document rather than parsing from the break point. Here, the information extraction involves five phases. Initially, the sentence is split based on some order. Tokens are identified. The sentence is then properly framed. Parsing has been carried out based on the grammatical structures. This is followed by pattern matching based on the relations. The parse tree language usually consist of four conditions namely tree pattern, link condition, proximity condition and the return expression. The linked condition contains some predicate expression. The query evaluations are done using relational data base management system. Here, the sentences are split according to the noun and adverbs present in the sentence.

3.7. Handling user queries:

In order to observe similarities between data records, Weifeng Sue et al. (2012), proposed an approach called CTVS. This approach deals with the process of combining tags. The queries are supplied by the user. The resultant web page is generated based on the user query and is provided as input to the tag tree construction. The resultant tree enters into the data region identification. This area which is further split into many data regions enters through the record segmentation where the particular records are identified before merging of data. Then it enters into the query result section identifications. Finally, the query result records are identified.

Conclusion:

The study reveals about various techniques involved in data extraction from the web pages. The technique mainly focuses on memory management. Besides, the amount of time consumed has been significantly reduced. It has been observed that the extraction of data is made simpler. Besides, automation plays a significant role in improving the performance.

References


Yanhong Zhai and Bing Liu, 2006. “Structured Data Extraction from the Web Based on Partial Tree Alignment,” IEEE Transactions on Knowledge and Data Engineering, 18: 12.

