International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Novel Web Data Extraction Using Template Extraction and Filtering Non Information

Jaishree G Waghmare[1], Vikas B Maral[2]

[1] Savitribai Phule Pune University, K J College of Engineering, Kondhwa, 411048, Pune, Maharashtra, India

[2] Savitribai Phule Pune University, K J College of Engineering, Kondhwa, 411048, Pune, Maharashtra, India

Abstract: ***Web is huge repository of information which contains different types of data in various forms. As we need to extract only the relevant data from web. Web data extractors are used to automatically extract the data from web documents. To study the problems related to web data Extraction different scientific tools are used which has broad range of applications. As we want only relevant data is to be extracted from the web. In our proposed system data is extracted using template extraction. Template matching will be based upon depth and data similarity and also removing the non-information part from the web pages by using filtering. The proposed system works on input document of variable depth.***

Keywords: Information Filtering, Non Information, Template Extraction Unsupervised learning, Web data extraction

1. Introduction

The World Wide Web contains huge amount of information. The information is available in WWW is of various form like text, images, video and other multimedia components. But all of this information is not of user's interest. Users are only interested in relevant information of their choice. So to extract relevant data some mechanism or tools are required. The web data extractors are used to extract the relevant data from web documents which are based upon extraction rules. It automatically extracts data from a website. Web data extraction can be done by using three techniques called as manual technique, Supervised & Unsupervised. Manual programs are written called as wrappers & it uses some in built rules to extract data. The problem is that labor intensive task, expensive & impractical. User has to give sample of the data to be extracted in supervised technique [1] The problem with this technique is that to generate the rules manual labeling of training examples is required. Labeling the training example is time consuming. In unsupervised approach it automatically learns extraction rule and extract as much data as it can. Unsupervised techniques are Road-Runner, ExAlg and FiVa Tech. Road Runner use automatically generated wrappers to extract the data also it uses partial rule. Wrapper generation process depends upon similarity & differences of web pages. EXALG extract the structured data from a collection of web pages generated from the common template. The algorithm used by EXALG is Equivalence class generation .Fiva Tech is a page-level web data extraction which automatically detects the schema of a Website.FivaTech uses two learning algorithm as Tree merging and schema detection.

One of the unsupervised techniques called as Trinity. This technique works on the set of web a document and learns the extraction rule from set of web documents. It partitions the input document into three parts as prefix, separators and suffix. From this trinary is generated which is then traversed to create regular expression. So we propose a system in which we are combing features of trinary tree along with filtration of non information. We take multiple URLs from different site as an input document. After processing these URLs it will generate DOM tree for each one. Then it is going to show extracted data. We are also going to evaluate filtering techniques to create template extraction pattern to ignore irrelevant data.

2. Literature Review

As Internet is a big source of information. The whole data is useful to us if only the data is in the well-formed but if it is not then for extracting these kinds of data web data extractors are used. There are many approaches for extracting data from web pages. Automatically extraction of data from these pages is very important.

By studying different research papers, journals related to this topic, provided lot of useful information. This information consist of how the data is extracted from web pages and the different techniques used for web data extraction for supervised, unsupervised as well as semi supervised approach. Here are some papers that we have referred for making literature survey.

[1]The paper discusses an overview of the web data extraction techniques and compares all these techniques based on methods for web data extraction which are based upon automatic web data extraction.

[2] Describes the design and implementation of the Web Wrapper which has been based on pre-defined schema. It proposes web wrapper which can accurately extract the data from the Web source.

[3]Introduce novel algorithms for extracting templates from a large number of web documents from heterogeneous templates. It considers Web documents as a matrix and finds clusters with the matrix. Clusters are formed based upon similarity template structures in the documents so that the template for each cluster is extracted simultaneously.

[4] Extract structured data from deep Web pages using novel vision-based approach. It uses visual features of deep Web pages. It carry out using mainly four steps like Visual Block

tree building, data record extraction, data item extraction and visual wrapper generation and also capture the amount of human effort needed to produce perfect extraction.

[5]Introduce new technique and approach using programming languages like PHP and AJAX, MySQL for user interface. This paper proposes a system called Xtractorz which perform web data extraction in a Mashup format. As the data cannot be directly extracted into a new structured form.

[6]Studies the problem of extracting data from a Webpage that contains several structured data records as well as propose a novel partial alignment technique based on tree matching. This approach enables very accurate alignment of multiple data records. Proposed technique is able to segment data records, align and extract data from Web very accurately.

[7]Proposes new Web data extraction approach called as FiVaTech is introduced which covers the problem of page-level data extraction. The page generation model is formulated using an encoding scheme based on tree templates and schema. They proposed page generation model with tree-based template matches the nature of the Web pages.

[8]proposes if starting & ending source code of web pages is removed then it is going to reduce noise from web pages because main content of any news is at middle of web page.

[9]It discusses different partially supervised web page cleaning techniques like segmentation based, template based, classification based, SST based cleaning technique

[10] It uses Case-Based Reasoning approach to find noise pattern. Noise patterns are classified using back propagation neural network algorithm.

[11] It introduces a new tree structure called as featured DOM tree. It focuses on detecting & eliminating local noise from web pages.

3. Proposed System

In our proposed system as shown in fig 1.we are using template extraction for extracting relevant data and also filtering is used to remove irrelevant information from the web page. We are using number of URL i.e. three URL`s from different sites as input document then HTML parser is applied on downloaded input to create DOM tree for each of URL. To remove non information to get relevant data filters are applied to generate modified DOM tree. Modified DOM tree is traversed using DFS traversal. Then trinary tree is generated on which node analysis is performed to generate final extracted data. From these different websites will find out common data then will perform the template matching which will based upon data similarity.

A similarity measure is a function which computes the degree of similarity between a pair of vectors or documents, a similarity measure can represent the similarity between two documents, two queries, or one document and one query • There are a large number of similarity measures proposed in the literature. The Data similarity The Data similarity used in proposed system are Jaccard coefficient, Dice coefficient, Canberra and Euclidean.

Jaccard coefficient is mainly used to similarity and diversity of sample sets. It is given by size of intersection divided by size of union. The Sorenson or Dice Coefficient is used for comparing similarity of two sample sets.

The Canberra measure is used to find distance between two points in vector. The Euclidean is used to find straight distance between two points.

Our proposed systems will use following filtering methods as

Tag based filtering, Data Type based Filtering, and Tag Tree based Filtering.
Tag based Filtering method works on filtering common data tags that are of no use or those will not have required relevant data. For e.g. header records, repeated menu, templates, copyright related text.

Data Type based Filtering technique will work on type of data that is expected to be extracted for further processing, such as images, links, dates, events.

Tag Tree based Filtering techniques partitions page into several content blocks which is based on some selected tags in the markup tag tree. Based on a set of data and tree structure it identifies the tree nodes or part of page that will have relevant required data. Then each node data is measures for usability with respect to topic of data that needs to extract from tag tree and rest of nodes are eliminated for further processing.

Algorithm

Step 1: Download the multiple URL as a input.
Step 2: Parse the input using HTML parser.
Step 3: Create DOM tree.
Step 4: Create Modified DOM tree by applying filters on DOM tree.
Step 5: DFS Traversal is applied on Modified DOM tree.
Step 6: Trinary Tree is created.
Step 7: Perform node analysis classifying it into template nodes and data nodes.
Step 8: Template is extracted i.e. final extracted data

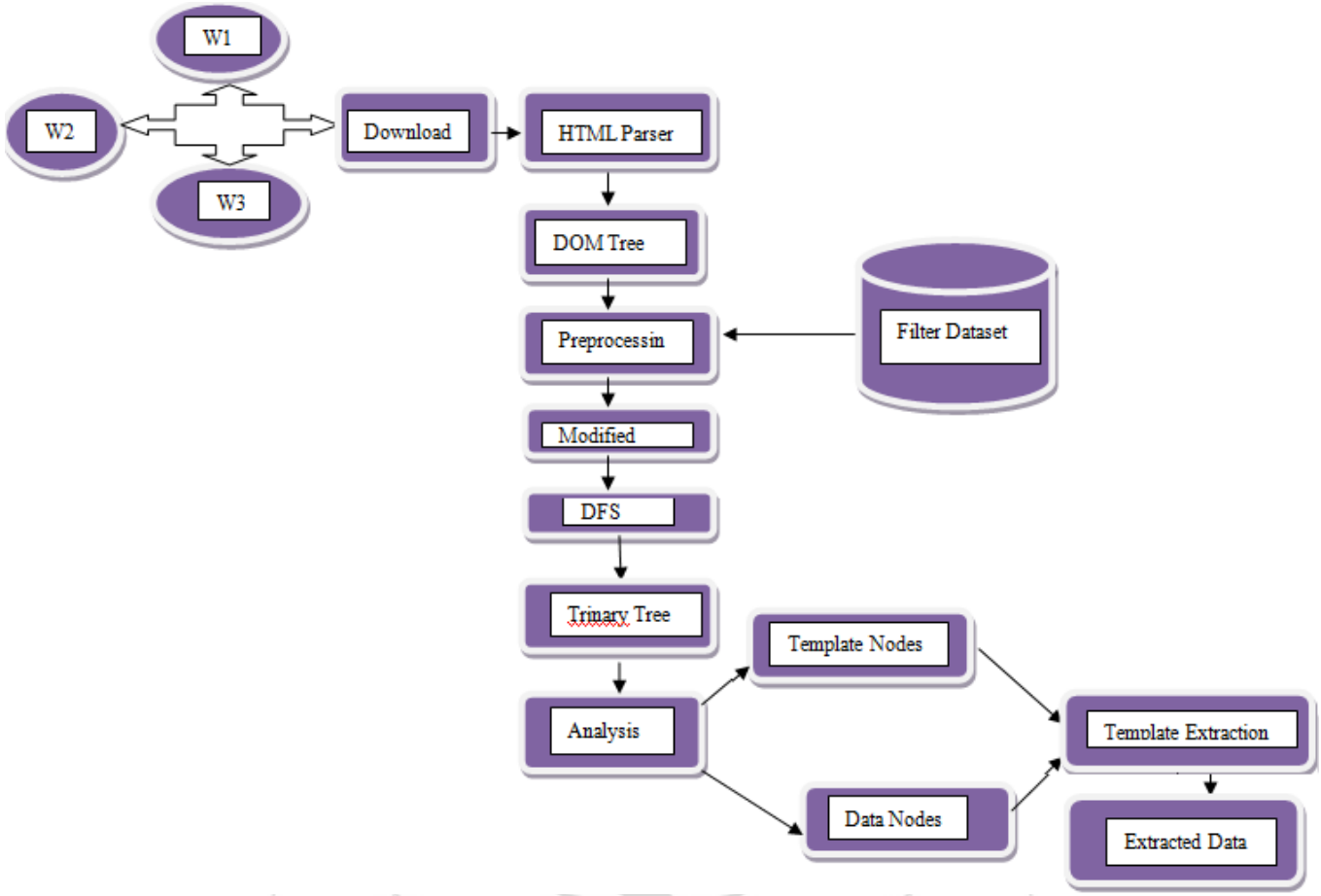

Figure1: Proposed Architecture

4. Conclusion and Future Scope

We presented web data extraction approach using template extraction that will extract variable length data. Also we are using preprocessing to remove non information part that is evaluating filtering techniques to create template extraction pattern to ignore irrelevant data. As the web documents contains large number of data with lots of irrelevant data. So from this huge information we need to filter only the relevant data that can be used for further processing. The problem with existing system is that it cannot handle the input document with variable length as well as filtering part is not present in it. By using our approach we are trying to create technique that will handle variable length data rather than fixed data length.

This system can be further enhanced by making system handled any dynamic data contents with creating patterns dynamically so that system or technique will learn and create or modify patterns dynamically. It will require less or rminimal human intervention after some time as system will start recognizing pattern dynamically

References

[1] Vidya V L, "A Survey of Web Data Extraction Techniques", International Journal of advance research in computer science and management studies, vol. 2, Issue 9, Sep. 2014.

[2] Suzhi Zhang, Peizhong Shi, "An Efficient Wrapper for Web Data Extraction and its Application", Proceedings of 2009 4th International Conference on Computer Science & Education.

[3] Chulyun Kim, Kyuseok Shim "TEXT: Automatic Template Extraction from Heterogeneous Web Pages", IEEE Transaction on knowledge and Data Engineering, Vol.23, No. 4, April 2011.

[4] Wei Liu, XiaofengMeng, Weiyi Meng, "ViDE: A Vision-based Approach for Deep Web Data Extraction", IEEE Transaction On knowledge and Data Engineering, Vol.22, No.3, March 2010.

[5] Rudy AG.Gultom, Riri Fitri Sari, Bagio Budiardjo, "Implementing Web Data Extraction & Making Mashup with Xtractorz", IEEE 2nd International Advance Computing Conference, 2010.

[6] Yanhong Zhai, Bing Liu," Web Data Extraction Based on Partial tree alignment", ACM 1-59593-046-9/05/0005.

[7] Yanhong Zhai, Mohammed Kayed, Chia-Hui Chang "FiVaTech: Page-Level Web Data Extraction", IEEE Transaction on knowledge on Data Engineering,Vol.22, No. 2, Feb 2010.

[8] Hu Fei, Li Ming, Ma Yan" Eliminating Noisy Information in Web Pages based on Source Code Shrinking*", International Journal of Advancements in Computing Technology (IJACT), Vol.4, No. 18, October 2012.

[9] S.S.Bhamare, Dr. B.V.Pawar" Survey on Web Page Noise Cleaning for Web Mining" International Journal

of Computer Science and Information Technologies, Vol. 4 (6), 2013.

[10] Thanda Htwe" Cleaning Various Noise Patterns in Web Pages for Web Data Extraction" International Journal of Network and Mobile Technologies, ISSN 1832-6758,Vol.1,Issuue 2, Nov 2010.

[11] Alpa K. Oza, Shailendra Mishra" Elimination of Noisy Information from Web Pages" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Vol. 2, Issue-1, March 2013.