Effective and Enhanced method for Template Extraction from Heterogeneous Web Pages

P. Rajeswari, A. Gandhirajan and R.Senthil

* Department of Computer science, Marudupandiyar College, Tamilnadu, India

Abstract- To achieve high productivity publishing the web pages are automatically evaluated using common templates with contents. The templates provide readers easy access to the contents guided by consistent structures. Cluster the web documents based on the similarity of underlying template structures in the documents so that the template for each cluster is extracted simultaneously. This process proposes to represent the document and a template as a set of paths in a DOM (Data Object Model) tree. As validated by the most popular XML query language XPATH, paths are sufficient to express tree structures and useful to be queried. Our experimental results with real-life data sets confirm the effectiveness and robustness of our algorithm compared to the state of the art for template detection algorithms.

Index Terms- Template extraction; clustering; minimum description length.

I. INTRODUCTION

The World Wide Web (WWW) is widely used to publish and access information on the Internet. In this paper, we extract template from these heterogeneous templates using text clustering. In order to achieve high productivity of publishing, the web pages in many websites are automatically populated by using common templates with contents. For human beings, the templates provide readers easy access to the contents guided by consistent structures even though the templates are not explicitly announced. However, for machines, the unknown templates are considered harmful because they degrade the accuracy and performance due to the irrelevant terms in templates. Thus, template detection and extraction techniques have received a lot of attention recently to improve the performance of web applications, such as data integration, search engines, classification of web documents.

II. ALGORITHM

In this paper, in order to relieve the limitations of the state-of-the-art technologies, this process investigates the problem of detecting the templates from heterogeneous web pages. This process proposes to represent the document and a template as a set of paths in a DOM tree. As validated by the most popular XML query language XPATH, paths are sufficient to express tree structures and useful to be queried. By considering only paths, the overhead to measure the similarity between these documents becomes small without significant loss

III. RELATED WORK

Template extraction from heterogeneous web pages is categorized into two areas; the first area is the site-level template detection where the template is decided based on several pages from the same site. Crescenzi et al. studied initially the data extraction problem in which the roadrunner extracts data template by comparing web page pairs. One page is considered as initial template, and the other page is compared with the template, which is updated when there are mismatches. Rajagopalan introduced the template detection problem. Previously, only tags were considered to find templates but Arasu and Garcia-Molina observed that any word can be a part of the template or contents. Vieira et al suggested an algorithm considering documents as trees but the operations on trees are usually too expensive to be applied to a large number of documents. Zhao et al. concentrated on the problem of extracting result records from search engines. For XML documents, Garofalakis et al. solved the problem of DTD (Document Type Descriptors) extraction from multiple XML documents. While HTML documents are semi structured, XML documents are well structured, and all the tags are always a part of a template.

The other area is the page-level template detection where the template is computed within a single document. Lerman et al. proposed systems to identify data records in a document and extract data items from them. Zhai and Liu proposed an algorithm to extract a template using not only structural information, but also visual layout information.

IV. METHODLOGY

Algorithm Required
Algorithm: Min-Hash
Input: Web Pages
1) GetBestPair(Clusters, Documents)
1.1) initial C={cluster1,cluster2….documentN}
1.2) for each pair clusterI,clusterJ of Clusters in C
1.3) min MDLCost=0
1.4) MDLCost=calculate MDLCost(clusterI, clusterJ)
If (min MDLCost> MDLCost)
min MDLCost==MDLCost;
Store pair(clusterI , clusterJ);
1.5) cluster pages which having less MDLCost than other pair
1.6) update Cluster Set C by merging best pair in one cluster.
Parsing these web documents into an xml document using DOM model. This saves the time to find out best templates from large no of web document and also save the memory.

V. ADVANTAGES

It is scalable to a huge number of sites due to the automatic process. . In this paper, it presented by algorithms for extracting templates from a large number of web documents which are produced from heterogeneous templates .This algorithm provides better performance compared to previous algorithms in terms of space and time. Our technique consists of two steps: Identifying data records without extracting each data field in the data records; Aligning corresponding data fields from multiple data records to extract data from data records, to put in a database table. This process proposed an enhanced method based on visual information for step (1), which significantly improves the accuracy of our previous algorithm. For step (2), this process proposed a novel partial tree alignment technique to align corresponding data fields of multiple data records. Empirical results using a large number of this process pages show that the new two-step technique can segment data records and extract data from them very accurately.

VI. CONCLUSION

To represent the heterogeneous information, a new approach used for template detection. This process employed the MDL (Minimum Description Length) principle to manage the unknown number of clusters and to select good partitioning from all possible partitions of documents, and then, introduced our extended MinHash technique to speed up the clustering process. This process proposed a new approach to extract structured data from this process pages. Our method only requires page contains, which is almost always true for pages with data records.

REFERENCES

[1] Document Object Model (dom) Level 1 Specification Version 1.0, http://www.w3.org/TR/REC-DOM-Level-1, 2010.

[2] Xpath Specification, http://www.w3.org/TR/xpath, 2010.

[3] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD, 2003

[4] Z. Bar-Yossef and S. Rajagopalan, “Template Detection via Data Mining and Its Applications,” Proc. 11th Int’l Conf. World Wide This process (WWW), 2002.

[5] V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites," Proc. 27th Int'l Conf. Very Large Data Bases (VLDB), 2001.

[6] M.N. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri, and K. Shim, "Xtract: A System for Extracting Document Type Descriptors from Xml Documents," Proc. ACM SIGMOD, 2000

[7] K. Lerman, L. Getoor, S. Minton, and C. Knoblock, "Using the Structure of Web Sites for Automatic Segmentation of Tables," Proc. ACM SIGMOD, 2004

[8] K. Vieira, A.S. da Silva, N. Pinto, E.S. de Moura, J.M.B. Cavalcanti, and J. Freire, "A Fast and Robust Method for Web Page Template Detection and Removal," Proc. 15th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2006.

[9] Y. Zhai and B. Liu, "Web Data Extraction Based on Partial Tree Alignment," Proc. 14th Int'l Conf. World Wide Web (WWW), 2005.

[10] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully Automatic Wrapper Generation for Search Engines," Proc. 14th Int'l Conf. World Wide Web (WWW), 2005.

[11] H. Zhao, W. Meng, and C. Yu, "Automatic Extraction of Dynamic Record Sections from Search Engine Result Pages," Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB), 2006.

AUTHORS

First Author – P.Rajeswari, Department of Computer science, Marudupandiyar College, Thanjavur, Tamilnadu, India.
Second Author – A. Gandhirajan , Department of Computer science, Marudupandiyar College, Thanjavur, Tamilnadu, India.
Third Author – R.Senthil, M.Sc., Ph.D, Department of Computer science, Marudupandiyar College, Thanjavur, Tamilnadu, India.