International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-6 December-2012

Template Extraction from Heterogeneous Web Pages

Trupti B. Mane[1], Girish P. Potdar[2]
Pune Institute of Computer Technology, Pune, India
B.E.Computer[1], M.E. Computer[2]

Abstract

The World Wide Web (WWW) is getting a lot of attention as it is becoming huge repository of information. A web page gets deployed on website by its web template system. Those templates can be used by any individual or organization to set up their website. Also the templates provide its readers the ease of access to the contents guided by consistent structures. Hence the template detection techniques are emerging as Web Templates are becoming more and more important. Earlier systems consider all documents are guaranteed to conform to a common template and hence template extraction is done with those assumptions. However it is not feasible in real application. Our focus is on extracting templates from heterogeneous web pages. But due to large variety of web documents, there is a need to manage unknown number of templates. This can be achieved by clustering web documents by selecting a good partition method. The correctness of extracted templates depending on quality of clustering.

Keywords

Template extraction, Clustering, Data mining, Information search and retrieval.

1. Introduction

The World Wide Web (WWW) is getting a lot of attention as it is becoming huge repository of information. The information is in the form of structured data and unstructured data. Many web pages get deployed by using common template. Web readers easily access to contents by using templates. More and more people are interested to create their own websites, for this reason there have been several different ways to ease out the procedure of website making and its maintenance. One of the many ways of an easy solution for making a website is the use of Website Templates. These templates offer the use of the various different designs and graphics which are already coded in the HTML format. Web templates reduce or eliminate professional web designers. Web template lacks only content and photos. Web templates can be used by any individual or organization to set up their website. Once a template is purchased or downloaded, the user will replace all generic information included in the web template with their own personal, organizational or product information. Templates can be used to effective separation between presentation logic and business logic. The unknown templates are considered harmful because they degrade accuracy and performance due to irrelevant terms in template. Now days, template detection and extraction have lot of attention to improve the performance of web application such as data integration, search engines, classification of document etc. A good template extraction technology can significantly improve the performance of such application. Fully automatic wrapper generation for search engines [1], presents method which automatically producing wrappers that can be used to extract search result records from dynamically generated result pages returned by search engines. The problem of extracting templates which conform to common template has been studied in [2], [3], [4]. Due to the assumption of all documents being generated from a single common template, solutions for this problem are applicable only when all documents are guaranteed to conform to a common template. However in real application it is not feasible to crawl large number of documents and to classify them into homogeneous partitions in order to use these techniques. If we group web documents by using URL, there may be different appearance of pages. Hence we cannot group web documents by using URL.

To overcome the limitation of the above techniques, we have to develop such method, which will extract template from heterogeneous web pages. But due to large variety of web documents, there is a need to manage unknown number of templates. This can be achieved by clustering web documents by selecting a good partition method. The correctness of extracted templates depends on the quality of clustering. In this paper, we introduce an approach to extract templates from heterogeneous web documents. In section 2, we present a survey of related work. In section 3, we present an approach for template extraction. In section 4, we state our conclusion.

2. Related work

In recent years, many researchers have tried to improve performance of template detection methodology. Because template detection improve the performance of web application.

2.1 Data Extraction from HTML document

HTML document can be represented with a Document Object Model (DOM) tree, web documents are considered as trees. In Automatic Web News Extraction Using Tree Edit Distance [5], presents a domain oriented approach to web data extraction. This approach is based on highly efficient tree structure analysis. Here, they evaluate structural similarity between HTML pages and based on that, grouping of pages is done to form cluster and find generic representation of structure of pages within a cluster. Structure of web page can be described by a tree. Tree-edit distance is used to evaluate the structural similarity between pages. Restrictive Top Down Mapping (RTDM) algorithm is used to identify relevant text. But the worst case complexity of RTDM algorithm is $O(n_1n_2)$ where n_1 and n_2 are sizes of the two trees, but it performs much better than traditional top down mapping.

In Extracting structured data from web pages [2], extraction of data is done in 2 steps: 1. formally define a template and propose a model that describes how values are encoded into pages using a template 2. Present algorithm that takes as input a set of template generated pages; deduce the unknown template used to generate pages and extracts as output. However, if this approach is used for crawling, indexing, then there will be problem of automatically locating collection of structured pages. Roadrunner: Towards Automatic Data Extraction from Large Web Sites [3], introduced extracting data from HTML sites through the use of automatically generated wrappers. This paper develops a novel technique to compare HTML pages and generate a wrapper based on similarity and differences. Goal is automatic generation of wrapper that is without any prior knowledge of target pages and human interaction. Matching technique is used to compare the HTML codes of two pages and to infer a common structure and a wrapper.

A Fast and Robust Method for Web Page Template Detection and Removal [4], RTDM-TD algorithm is used to find optimal mappings between the Document Object Model (DOM) trees of web pages. This algorithm is based on a restricted formulation of top down mapping between two trees, which is particularly suitable for detecting structural similarities among web pages. But the operations related to trees are expensive.

2.2 Data extraction from XML document

XTract [6], provides a system for extracting Document Type Descriptor (DTD) from XML documents. XML document can be accompanied by a Document Type Descriptor (DTD) which plays the role of a schema for an XML data collection. DTD contains valuable information on the structure of document. XTract method solved the problem of DTD extraction from multiple XML documents.

2.3 Clustering method

In Automatic Web News Extraction Using Tree Edit Distance [5], presented a method, in which small number of sampled documents are clustered first, and then the other documents are classified to the closest clusters. In this approach selecting proper training data is not easy task. In Joint Optimization of Wrapper Generation and Template Detection [7], labelled training data is used for clustering.

3. Proposed System

Template extraction method consist of following steps:

- HTML document and Document Object Model
- Essential paths of document
- Template of document
- Representation of clustering
- Minimum Description Length
- Clustering algorithm

System S, for template extraction can be defined as
$S = \{W, T, C, DOM, P_W, S_P, E_P, t_{w,}\}$

- W is set of web document, W= {w1, w2, ..}.
- T is set of template, T = {t1, t2, t3, ..}.
- C is set of cluster, C = {c1, c2, ..}.
- DOM is Document Object Model tree, DOM = {d1, d2, ..}, for all w_i, there exist d_i.
- P_W is set of all paths in W.
- S_P is support of path
- E_P is set of essential paths
- t_w is minimum support threshold.
- Function f, f (T-> W), one to many.
- Function g, g (C-> W), one to many.
- Function h, h (T-> C), one to one.

For web document, using HTML parser, we can parse the web document and build the Document Object Model (DOM) tree. From DOM tree of web documents we can find out paths, there are two types

of paths content path and template path. Template paths contain only structured information; it does not contain actual contents. From DOM tree one can find essential paths, which represent template. The system model is shown in figure 1.

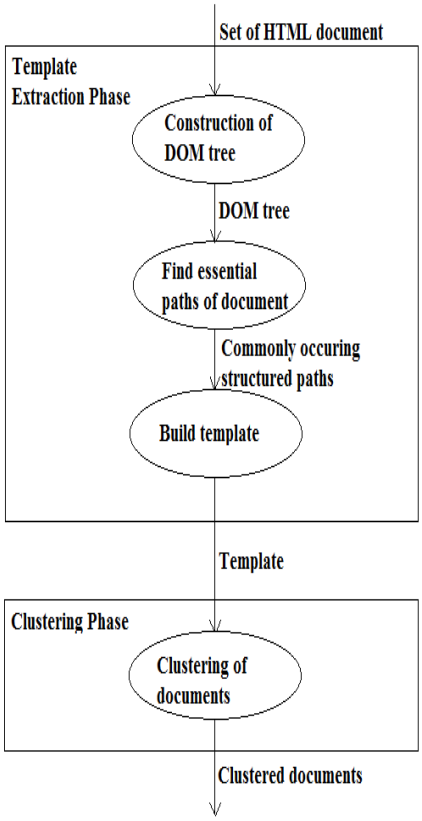

Figure 1: System model for template extraction

3.1 HTML Document and Document Object Model

HTML document can be represented by Document Object Model (DOM) tree. DOM is easy way to represent structured information of HTML document. DOM is application programming interface for valid HTML. It defines logical structure of document. DOM is platform and language neutral interface that will allow programs scripts to dynamically access and update the content, structure, and style of document. The HTML document, HTML element, texts in the HTML document, HTML attribute and comments are represented in DOM tree as document node, an element node, text nodes, attribute node and comment node respectively.

3.2 Essential paths of document

Define path set P_W for document W. P_W is set of all paths in W. Support of path(S_P) is the number of documents in W, which contain the path. For all w_i, there exists t_{wi}. t_{wi} is decided by taking mode of support values. If a path is contained by a document w_i and support of path is at least the given t_{wi} then the path is essential path of w_i. Set of such essential paths are E_P. These essential paths are used in extracting template. Set of paths and HTML documents are denoted by matrix of size $|P_W| \times |W|$

```
<html>
<head>
<title>DOM tuto</title>
</head>
<body>
<h1>DOM lesson</h1>
<p>Hello</p>
</body>
</html>
```

Figure 2: HTML Document

3.3 Template of document

A template is set of common layout and format feature that appear in a set of HTML pages that is produced by a single program or script that dynamically generates the HTML page content. Template of a document cluster is a set of paths which commonly appear in the documents of the cluster.

3.4 Representation of Clustering

Cluster c_i-> (T_i, W_i), T_i is set of paths representing the template of c_i and W_i is set of documents belonging to c_i. Successful condition for clustering is $W_i \cap W_j = \phi$ and $U_{1<=i<=m} W_i = W$

3.5 Minimum Description Length (MDL) principles

MDL principle is used to manage unknown number of clusters and to select good partitioning from all possible partitions of HTML documents. The MDL principle states that the best model inferred from a given set of data is the one which minimizes the sum of 1) the length of the model, in bits, and 2) the length of encoding of the data, in bits, when described with the help of the model.

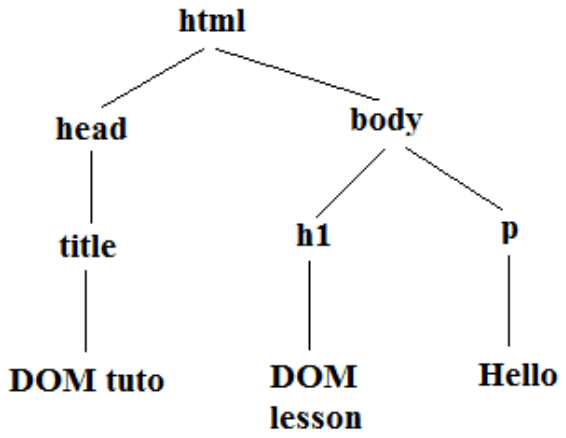

Figure.3: DOM tree

3.6 Clustering using MDL cost

Clustering is a process of forming groups (clusters) of similar objects from a given set of inputs. Good clusters have the characteristic that objects belonging to the same cluster are "similar" to each other, while objects from two different clusters are "dissimilar". Input to clustering is set of web document W= {w1, w2, w3, ..} and output is set of cluster C= {c1, c2, ..}. Cluster c_i-> (T_i, W_i), as explained in section 3.4. Clustering of web document can be done by using agglomerative hierarchical clustering algorithm as explained in [8]. Initially each document is an individual cluster. When a pair of cluster is merged, the MDL cost of the clustering model can be increased or decreased. Find a pair of cluster whose reduction of MDL cost is increased in each step of merging and the pair is repeatedly merged until any reduction is not possible.

Proposed system includes above six steps. System will take input as HTML document and automatically extract template within very short period of time as compared to manual method. Here, we describe psedo code for our system:

1. Procedure: START(W)
2. While $W \neq \phi$
3. CreateDOMtree(W)
4. FindPath(DOM)
5. FindSupporOfPath(DOM)
6. FindEssentialPath(P_W, S_P)
7. ExtractTemplate(E_P)
8. ClusterDocuments(W,T)
    a. GetMDLcost(c_i, c_j, C)
    b. GetBestPair(C)
9. End while
10. End procedure

4. Conclusion

Template extraction from heterogeneous web pages can be done by constructing Document Object Model (DOM) tree of HTML document and finding essential paths of document. Clustering of web document can be done on the basis of template structure to manage unknown number of template.

References

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

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

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

[4] 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.

[5] M. de Castro Reis, P. B. Golgher, A.S. da Silva, and A. H. Laender, "Automatic Web News Extraction Using Tree Edit Distance," Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.

[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] S. Zheng, D. Wu, R. Song and J.-R. Wen, "Joint Optimization of Wrapper Generation and Template Detection," Proc. ACM SIGKDD, 2007.

[8] Chulyun Kim and Kyuseok Shim, "TEXT: Automatic Template Extraction from Heterogeneous Web Pages" IEEE Knowledge and Data Engineering, vol. 23, no. 4, 2011.

Trupti Mane has received B.E. (Computer Science) degree in 2010 from Shivaji University. She is currently appearing for Master of Computer Engineering in Pune Institute of Computer Technology, Pune. Her research area is Data mining.