Comparative Cost Analysis of Template Extraction from Heterogeneous Web Documents

Jyoti Mhaske
Computer Department,
MET BKC Adgaon, Nashik, Savitribai Phule Pune University, Maharashtra India

ABSTRACT
Extracting structured information from unstructured and semi-structured machine-readable documents automatically plays vital role in now a days. So most websites are using common templates with contents to population the information to achieve good publishing productivity. Where Internet is the major resource for extracting the information. In recent days Template detection technique received lot of concentration to improve in different aspects like performance of search engine , clustering and classification of web documents , as templates degrade the performance and accuracy of web application for a machines because of irrelevant template terms. So Novel algorithms is useful for extracting templates from a large number of web documents which are generated from heterogeneous templates. Using the similarity of underlying template structures in the document cluster the web documents so that template for each cluster is extracted simultaneously.

Keyword
web Template extraction, clustering documents, minimum description length principle.

1. INTRODUCTION
World Wide Web is the most useful source of information. For improvement of productivity of publishing the WebPages in many websites are automatically populated by using the common templates with contents. The simple and unique templates provide easy access to readers to the contents directed by consistent structures. However for machines the unknown templates are considered harmful since they degrade the accuracy and performance of web applications due to irrelevant terms in templates[4][5]. Thus template detection techniques are play very important role to improve the performance of search engines web documents and classification of web documents.

The Web poses itself as the largest data repository ever available in the history of humankind[1]. Major efforts have been made in order to provide efficient access to relevant information from huge source of data. Although several techniques have been developed for Web data extraction, but their use is still not spread, mostly because of the need for high human intervention and the low quality of the extraction results. So here a domain oriented approach is used to Web data extraction and discuss its application to automatically extracting news from Web sites[1][4].

So here novel algorithms is discussed for extracting templates from a large number of web documents which are generated from heterogeneous templates and 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[1].

Grouping of web documents is not done on the basis of URL. In fig1 the pages look clearly different but their URLs are identical except the value of layout parameter. If by considering only URLs to group the pages then the pages from different cluster will be included in the same group[11][12].

**Fig1. Different template of the same URL**
Web document and its template are represent in figure 1 and a set of paths in a DOM tree are shown in table1, paths are used to express tree structures and also useful to be queried. By using only paths, overhead is occurs the similarity between documents becomes small without significant loss of information. For example, let us consider a HTML documents and paths in Fig. 2. Support rate of each tag is present in table1 and Paths are defined later. Document A is represented as a set of paths {p1; p2; p3; p4; p6} and the template of both A and B is another set of paths {p1; p2; p3; p4}.

**TABLE 1 Paths of Tokens and Their Supports (MDL minimum descriptor table)**

<table>
<thead>
<tr>
<th>ID</th>
<th>Path</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>Document(html)</td>
<td>4</td>
</tr>
<tr>
<td>p2</td>
<td>Document(html)(body)</td>
<td>4</td>
</tr>
<tr>
<td>p3</td>
<td>Document(html)(body)(h1)</td>
<td>3</td>
</tr>
<tr>
<td>p4</td>
<td>Document(html)(body)(br)</td>
<td>3</td>
</tr>
<tr>
<td>p5</td>
<td>Document(html)(body)\List</td>
<td>3</td>
</tr>
<tr>
<td>p6</td>
<td>Document(html)(body)(h1)\Tech</td>
<td>1</td>
</tr>
<tr>
<td>p7</td>
<td>Document(html)(body)(h1)\World</td>
<td>1</td>
</tr>
<tr>
<td>p8</td>
<td>Document(html)(body)(h1)\Local</td>
<td>1</td>
</tr>
</tbody>
</table>
2. EXISTING SYSTEM

Extracting well Structured Data from different web pages are used for improve performance of search engines classify web documents. Automatic Web Extraction Using Tree edit distance: By using DOM tree it does the cluster of the documents. Automatic Template Extraction from Heterogeneous Web Pages and clustering documents by using MDL algorithm[2].

Data extraction work can be classified along different dimensions: sources of information targeted, percentage of data or information that is in the form of automation, complexity of data extracted (flat vs. nested). The template extraction problem can be categorized into two levels or area. The first is the site-level template detection where the template is decided based on several pages from the same site. They detect elements of a template by the frequencies or occurrence of words , but we consider the MDL principle as well as the frequencies to decide templates from heterogeneous documents[7][8]. While HTML documents are semi structured and XML documents are very well structured, but all the tags of web documents are always a part of a template. The solutions for XML documents fully utilize these properties. In this problem of the template extraction from heterogeneous document how to do partition of given web documents into homogeneous subsets is important[3][4].

To overcome the limitation of techniques with the web documents, the method of extracting templates from heterogeneous web pages are carried out here. Generally webpages are represented by HTML documents. These web documents are considered as trees for clustering. Because of 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 generate from common template. However in real web applications, it is not feasible to classify frequently crawled documents into homogeneous partitions in order to use these techniques. Here the DOM methods of constructing trees are used which can be easy to handle larger number of web documents. This method like Tree edit distance is very expensive and high rate of time Complexity, so DOM construction is used for complexity reduction purpose[1].

3. PROPOSED SYSTEM

Here the architecture of the system in which it describes the entire system works process. Various heterogeneous web documents are collected and by splitting a tag of html documents, a tree is constructed from the paths specified in MDL table. It all depends on the similarities of the documents and the paths, clustering process is done. Various clustering techniques are used and cost is calculated for each clustering technique. In the MDL clustering technique taking each document as individual cluster, pair of clusters are merged in order to as reduce the final cost. Here different web pages are providing as input to system. Each web page has different or may also same document structure. After that parse the web documents into an xml document using DOM model[2]. After that find out the paths using its tag entry in the XML document. After that applying the MinHash algorithm is used to find out best pair pages from given input Pages. Then classify these best pair pages into different groups. Recommendation is depends on these groups to the user. This saves the time to find out best templates from large no of web document and also save the memory i.e. need to find out the best template structure.

4. SYSTEM ARCHITECTURE

In the proposed system that extracts templates from web pages using Template Extraction from Web pages algorithm. The overall architecture of the proposed system is shown in fig.3.

More than single Web Pages which have different templates are downloaded. Downloaded Web Pages are parsed and constructed DOM tree in order to reduce into number of small blocks. These Blocks are made of both content and non content blocks. The templates stored in database can be further used by the web designer to develop Webpages which improve the performance of the search engines. It also able to the web designer to develop web pages as fast and easy. separate the non content blocks from the blocks of web pages. Some items in a non content block will reduce the performance of search engine. Hence hierarchical agglomerative clustering is used to find the common non content blocks among the web pages downloaded. This clustering technique is used to group the common items in the each non content blocks into a cluster. Cost efficiency of this algorithm is reduced by applying Minimum description Length principle by eliminating the clusters which uses large number of bits to define itself. Finally the Templates are extracted and stores in database system for further use.

The algorithm for template extraction is as follows:
1. Download Web pages which are to be composed[2].
2. Parse the Web pages using DOM parser and construct a DOM (Document Object Model) tree[2],[3].
3. Convert the DOM tree into visual Blocks using VIPS (Vision-Based Page Segmentation) algorithm.
4. Identify the non- content blocks.
5. Cluster the non-content blocks using Hierarchical Clustering method.

Fig 3: System architecture of Template Extraction from Heterogeneous Web Pages

5. METHODOLOGY

TEXT-HASH: It is the agglomerative clustering algorithm with MinHash signatures discussed in it requires an input parameter which is the length of MinHash signature.

TEXT-MAX: It is the clustering algorithm with both MinHash signatures and Heuristic 1 to reduce the search space. It requires the length of the signature as an input parameter.

Algorithm Required
Algorithm: Min-Hash
Input: Web Pages
1) GetBestPair(Clusters, Document)
   1.1) initial C={cluster1,cluster2,…,documentN}
   1.2) for each pair clusterL,clusterJ of Clusters in C
   1.3) min MDLCost=0
   1.4) MDLCost=calculate MDLCost(clusterL, clusterJ ) If
       (min MDLCost> MDLCost) min MDLCost==MDLCost;
       Store pair(clusterL, 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.

6. MODULES DESCRIPTION

1. Document Collection and DOM representation of web
documents: first collect the HTML documents as input. Then
DOM defines a standard for accessing documents, like
HTML and XML. The DOM is used to presents an HTML
document as in the form of tree structure.

2. Essential Paths and Matrix calculation in the given a web
document collection D ={d1,d2,…, dn}, then define a path
of web documents set PD as the set of all paths in D. Note
that since the document node is a virtual node shared by
every document, do not consider the path of the document
node in PD. The support path of documents is defined as the
number of documents in set of documents i.e. D, which
contains the path. So matrix calculation is used for cost
calculation of template extraction techniques of different
algorithm.

3. Agglomerative with MINHASH (TEXT-HASH) In this
system, although take only essential paths, the dimension of
Ei is still high and the number of documents is large. Thus
the complexity of TEXT-MDL is O(n^2) still expensive. In
order to avoid this situation the estimation of the MDL cost
of a clustering by MinHash not only to reduce the dimensions
of documents but also used to find the best pair that is used to
merged in the MinHash signature space.

7. CONCLUSION

The template detection and extraction techniques are used in
heterogeneous web pages. Cluster the documents based on
the template used in the web pages, also extract the data used
in the web pages. By using this web pages are fully studied
and their contents are compared and extracted. Although
extracting templates from heterogeneous web pages needs
large time to extract and detect. So time and cost can reduce
by using MinHash algorithm. From the proposed system
results, can predict that hyper graph technique is much
helpful for extracting templates from different web pages.
There is some future work as proposed work extended to
derive multiple templates and the extracted templates can be
used for the individuality and the templates with the same
details can be deleted.

8. REFERENCES

knowledge and data engineering, VOL. 23, NO. 4,APRIL 2011.


