Web Entities Extraction Based on Semi-Structured Semantic Database

Fang Dong, Mengchi Liu, and Kun Ma

State Key Lab of Software Engineering, School of computer, Wuhan University, Wuhan, China
Email: whdongfang@163.com, mengchi@cs.carleton.ca

Abstract—Web is the biggest source of information and contains many entities and relationships between them, extracting these data from Massive Web pages and Integrating to a Semi-Structured Data with rich semantics will be more conducive to the management and use of these web data. On this premise, a comprehensive method is proposed to perform extraction the entities and relationships from the webpages. The method consists of two steps: 1) The target Web pages which contains these entities will be found based on the combination of vision information and content of keyword, meanwhile recording the relationship between father and children target Web pages; 2) Extracting the entities with analysis of DOM tree structure of the obtained Web pages and definitions of some extraction rules. At last, the extracted data is organized into a Semi-Structured Data with special relationships. Experiments on a large number of HTML pages have showed that this method can get a high correct rate and coverage.

Index Terms—Web Entities, Data Extraction, Semi-Structured Semantic Database

I. INTRODUCTION

With the explosive growth of WWW information, more and more Web information needs to be extracted and reasonably effective managed. Web can be seen as a data source contains abundant information and a large number of data can be extracted according to different application background. Most of data extraction is concerned with a traditional database named a relational database. Because relational database is based on strict mathematical logic, and easy to use, in a fairly long time it is considered to be an ideal database model. But with the improved requirements of data management, database has changed to adapt to diversity of data structure such as MMDB[22], INMDB[23].To extract data information from web pages and store the information in a semi-structured database—INMDB. From the perspective of HTML page structure, the design of HTML pays attention to the display of the data rather than the data manipulation, so data organization on the web page which is not clear has caused great difficulties to web data integration. On the other hand, the goal data format of data extraction—INM is a model which can express complex and dynamic relationships. So its structure is much more complex than relational databases. Thus, massive data extracted from the disorderly pages and stored in the INM database is a complex and vast project.

The Web information extraction for a rich semantic of semi-structured data is divided into two parts: one is how to find the Web pages which contain entities, and the other is data extraction after finding the target Web pages. The first part is actually about the focused crawler, also called web robot or spider-a program that gathers automatically web pages by exploring the linked structure of the Web [1] [2]. Early focused crawlers use the domain keywords to determine if the page is relevant or not after downloading it [3]. Ontology has been used to detect the relevant score for links before downloading in the literature [4] [5] [6] and [7]. Some researches analysis the structure hyperlinks to evaluate the relevance. A focused crawler approach that evaluates the page content relevance using ontology and hyperlinks analysis has mentioned in the literature [8]. M. Jamali uses the link structure analysis with the similarity of the page context to determine the download pages priority [9] and the hyperlinks are helpful for discovering the relationships between the web pages [10]. After finding the target web pages, next step is to extracting data through analysis of web pages obtained. System ViDRE [11], ViNTs [12] and ViPERS [15] relay on visual information to extract data, MDR [13] and DEPTA [14] are two extracting system created from the Department of computer science in Chicago University. MDR locates and extracts data based on HTML DOM tree structure and DEPTA improves the MDR with adding visual information. SGWRAP [17], EXALG [18], XPath-Wrapper [19] and WDE [20] are the technology which is similar to MDR using Web page Positioning, and [16] is about extraction approaches for image retrieval.

But INM has established model for the real world through the description of relationships between objects and objects, objects and relations, relations and relations. It is a data model which can describe the complex relationships between the real world objects as while as context relations. So INM is more accurate and more natural for the real world modeling. While in general, HTML webpages have always organized data according to the real world. So the Information Network data model can be combined with structure of HTML webpage to
deduct the extraction rules. On this foundation, Web entities extraction method based on INM includes the following several key problems:

(1) How to get data page through crawler program, and meanwhile recording the relationship between web pages.

(2) How to collect the INM database data model, which is data structure obtained can be used to deduct the rule of data extraction.

(3) How to analyze the extraction rules according to the features of web pages and INM data model, and generate the corresponding the INM rules mode tree for data extraction.

This paper will focus on the key issues and discusses corresponding solutions. Chapter 2 has introduced data structure of INM and operating mechanism of this data extraction system which consists of 3 parts. Chapter 3 has explained the target page access and INM deserialization problems, this paper adopts the combination of keyword inspiration, content matching and HTML structure to implement focused crawl, and then INM has been deserialized for the next step of data extraction. Chapter 4 has illustrated that the data extraction rules are inserted into INM tree using visual feature and HTML feature, and the INM tree has changed to be a tree with extraction rules. Chapter 5 has proved the effectiveness of proposed method through university experiment. Chapter 6 summarizes the whole article and the next work step. This paper has proposed a method to extraction entities from the web for semi-structured database, first to acquire the target page with focused crawler and then to extract the data with extraction rules based on the first step. This paper has done a detailed study on web data extraction for the network of semi-structured database data, and use a combination of a variety of methods to extract the information in heterogeneous network data source. So it has certain positive meaning in data extraction for semi-structured database.

II. OVERVIEW OF THE INFORMATION NETWORK MODEL AND EXTRACTION SYSTEM

Currently, data extraction results are usually stored in a relational database, such as Oracle, My SQL. But the relational database has become increasingly difficult to adapt to the management of data and INM database modal is the reflection of real world. This article focused on the data extraction for INM database which is a Semi-Structured Semantic Database. So it is necessary to introduce the Information Network Model first and then to explain this data extraction mechanism.

A. Introduction of Information Network Model

Currently data extraction mainly faces to traditional databases, such as relational databases. However, the traditional relational database model may not describe complex semantic information, so there are a lot of new database models have been proposed, and Information Network Model is one of them. In the Information Network Model, there are the concepts of model, instance, relations, attributes, role relationships and object classes these concepts have been designed based on the characteristics of the entity itself or relationships between entities. Different relationships between entities are defined in the INM and different attributes describe the basic features of the entities. So INM can expression the complex, dynamic, multi-level contact between the natural. Fig 1 is a Part of University structure base on INM.

Create abstract class University. Teaching unit subsume disjoint{[College, Department, Lab subsume partition{[…, …]}]} [Faculty [specific Research Area (M: N): Research Area, role-based@Office Phone: string, @Start Time: date]]

(disjoint{

Professor [static (…, …)],
Associate Professor [static (…, …)],
Lecturer [static (…, …)],
Tutor [static (…, …)],
Postdoctor,
External Professor -> [Visiting Professor, Consulting Professor],
Senior Technical -> [Engineer, Senior Engineer, Primary Lab Technician,
Intermediate Lab Technician, Senior Lab Technician ]} (inverse Professional Title) (M: N),
Tutor -> {
Master Tutor,
Doctoral Tutor} (inverse Academic position*),

Figure 1. INM modeling language of teaching unit

The University contains multiple organizations and education unit is the most important one which contains most of the information, the URL in this unit Web pages point to the hyperlink which contains college and department data set of records and other information in the actual university. The University, teaching unit, college and department belong to the object class; faculty belongs to role relationship class.

The relation between University and teaching unit (college and department) belongs to contain relationship;

The college and the department are the subsume of teaching unit;

The relationship between faculty and some college or department is the role they play in these units, that is role-relationship;

This paper only discusses the entities extraction of object class and role relationship in the university and those are an important part of INM. The most important object class and role relationship class instance in university is college, department and faculty. Therefore, the aim of this paper is to extract colleges, departments and faculty data in the web pages and store them in the INM database. The extraction problem is to find out how many subsume the teaching units have and that is how many colleges and departments the teaching units contain, how many departments each college contains, how much faculty belongs to a college or department, and what the faculty’s title in these units.

B. Data Extraction Mechanism

The data source of this data extraction system is every university homepage, and the final extraction result will be stored in the INM database with INM data structure. Figure 2 shows the data extraction implementation mechanism.
In crawling subject page program of the colleges and departments, the URL on the university home page usually indicates that the anchor text links pointing to the college or department record collection page. The page s including both college and department record set usually contain some special keywords, such as colleges, departments, and the record set containing special keywords in the page are presented in the form of a regular in the HTML page. So this paper adopts three ways combined to get target Web pages:

**The anchor text of URL**: the current page of anchor text keywords is college or department;

**Web page relevance**: adopting the formula 1 to calculate the topic relevance, choose one of the biggest correlation web pages.

\[
sim(q, p) = \frac{\sum_{k \in q \cap p} f_{kp}}{\sqrt{\sum_{k \in q} f_{kp} \sum_{k \in p} f_{kp}}}
\]

q is subject; p is web pages crawled; \(f_{kp}\) is the frequency of term k in q; \(f_{kp}\) presents the frequency of term k in p.

**Visual features**: the tag in the page starting position (x, y) of the abscissa and ordinate full compliance, abscissa and ordinate consistent with equal distance, horizontal ordinate consistent with equal distance.

Faculty Webpage is special because the page does not contain some special keywords such as colleges and departments. So here to build a database of names which is Name.data, and page matching is one of the most record set is contained in the page.

### B. INM Model Collection

The INM communication between client and server has followed INMTP protocol (model of information networks communication protocol). As same as other application layer protocol, INMTP protocol message can be divided into the request message and response message, the transport layer adopts the TCP protocol. The client only needs to encapsulate INM database of DML (data manipulation language) and the DDL (database definition language), and their lengths to constitute the client's message. Client only constructs the statement of INM language and do not try to understand the INM language specific meaning.

In the concrete realization, the INM database model is divided into four structures to describe the class of objects, role relationship classes, relationships and attributes in the schema. After the server receipt the statement sent by the client, it can get the length of the statement with converting byte sequence, and then determine the scope of the statement, at last call server-side parser. The structure of return data is a tree structure, so it needs to be serialized and converted into a sequence of bytes stream. Module of mode is only responsible for model access and the response message format is fixed, the message type is considered the model query message and the serialized data result as a tree structure. Based on Serialization algorithm on the server, here is deserialization algorithm on client.

---

**Figure 2. Data extraction mechanism**

The system generally consists of three parts

Part1: Web pages will be access to through web crawler to prepare for the next page preprocessing. A non-standard HTML page must be converted to standard grammar HTML, then parse into a DOM tree structure based on HTML tag syntax rules.

Part2: To collect INMDB the model information. That is analyzing server response and giving information to the module of deserialization according to request to server with communication protocol, and then it can generate INM mode tree.

Part3: The feature of data form in HTML pages can be combined to INM pattern tree to create a pattern tree with extraction rules. Then wrapper program will automatically extract data in HTML pages and generate database instance, and store in the INM database.

### III. FOCUSED CRAWLING AND INM MODEL COLLECTION

#### A. Focused Crawler

To extract entities of college, department and faculty in a university Website, the first step is to obtain the Web pages contains this information. Figure 3 has shown that the different paths from the university homepage to target Web pages. Focused crawler program must detect the relevant Web pages for links before downloading.

In this picture, the solid is the part to be extracted and the dotted line is not related webpage and hyperlink. If starting from the home page of a university, when removing duplicate links. The whole university structure can be seen as a tree structure, and that is similar with the structure of INM corresponding university. But the faculty record set pages directly linked to the university homepage are not to be crawled because not knowing that the faculty belongs to which unit. In this multi-tree, only the web pages of colleges, departments and faculty are needed to be crawled for the next step of extraction. The data need to be extracted are signed in the following figure.
Figure 3. Special path of target Web pages

Algorithm 1. deserialization of INM

Input: the model of INM database
Output: the chain structure of the tree
Begin
1 Deserialize_result_node(& Sequence)
2 Type = Deserialize_get_byte(&Sequence);// The value type of the query return
3 Var = Deserialize_get_string(Sequence,&len);// variable name of the query return
4 Des = Deserialize_get_byte(Sequence,&len);// Descriptive string
5 If Type == RESULT_NODE_TYPE_STRING
6 Then Value = Deserialize_get_string(Sequence,&len);// Results the values of variables
7 Else If Type == RESULT_NODE_TYPE_STRUCT
8 Then Value = Deserialize_schema(Sequence);
9 Cmts = Deserialize_get_byte(Sequence,&len);// The query value remark
End

IV. ENTITIES EXTRACTION OF OBJECT CLASS AND ROLE RELATIONSHIP CLASS

Extraction rules can be formulated for the INM database model according to various features of web page. Then the extraction rules can be inserted into model tree to help extracting relevant data. After observing, we have found that most of web pages which contain INM entities in the university are shown in a certain form. This paper mainly makes the extraction rules with DOM tree feature, visual feature and semantic feature.

Definition 1 Separating tags:
All data are encapsulated in the HTML page tag. In the web pages which contain college and department record set the data to be extract are the names of college and department. While the names of college and department are separated from other data by HTML tags such as <h1>, <h2>, <h3>, <h4>, <p>, <tr>, <dt>, <ul>, <strong>, <img>, <a>. The table I react the distribution of school separate tags in 132 universities in INM database. Because the names of college and department are almost showed as a text is a hyperlink, the data to be extracted are separated by <a> tag.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;a&gt;</td>
<td>62.7%</td>
</tr>
<tr>
<td>&lt;h1&gt;</td>
<td>5.4%</td>
</tr>
<tr>
<td>&lt;h2&gt;</td>
<td>4.7%</td>
</tr>
<tr>
<td>&lt;h3&gt;</td>
<td>4.7%</td>
</tr>
<tr>
<td>&lt;h4&gt;</td>
<td>3.1%</td>
</tr>
<tr>
<td>&lt;p&gt;</td>
<td>1.6%</td>
</tr>
<tr>
<td>&lt;tr&gt;</td>
<td>2.3%</td>
</tr>
<tr>
<td>&lt;dt&gt;</td>
<td>1.6%</td>
</tr>
<tr>
<td>&lt;ul&gt;</td>
<td>3.9%</td>
</tr>
</tbody>
</table>

Definition 2 Record set:
The HTML tags encapsulating the entities of college and department have similar locations in the Web pages and have the same parent node in corresponding DOM tree. So the data items which are encapsulate by HTML
tags conform to the following two conditions of the data item represent as a collection of \( H = \{ h_1, \ldots, h_s \} \).

1) The DOM tree structure feature: The HTML tags have the same Xpath in DOM tree.
2) HTML webpage visual feature: label position where the starting position of the webpage \((x, y)\) uniformly in horizontal or vertical coordinates and the abscissa ordinate with equal distance is consistent, the ordinate is consistent with equal distance abscissa.

### B. Features of Extraction

![Image of DOM tree structure](image)

**Figure 4. The real university webpage and its DOM tree structure**

Figure 4 is a real Web page which contains college entities and corresponding HTML DOM tree structure. As showed in the picture, the dotted line region is the data region of school entities and in the solid area is the data to be extracted.

Entities of object class and role relationship class in INM database are presented in the form of comparison rules. This paper has summed up some extraction rules through the observation of a large number of real pages which can help the work of data extraction. These features can be divided into DOM features and visual features.

**DOM features**
- D-Feature1: The same class of data item in the same web page has identical Xpath.
- D-Feature2: separate label of extraction data are \(<h1>\), \(<h2>\), \(<h3>\), \(<h4>\), \(<p>\), \(<strong>\), \(<img>\), \(<tr>\), \(<dt>\), \(<ul>\), \(<a>\).

**Visual features**
- V-Feature1: The position \((x, y)\) and font of the label element of the same data in a webpage are similar.
- V-Feature2: the extracted data usually occupy a large area and in the center of the webpage.

**Semantic feature**
- S-Feature: the name of object class always contains some special words, such as university, school and department.

### C. Design of INM Extraction Template

**Definition 3** INM model rule tree:

INM model rule tree is an INM model tree in substance, and its node domain contains information extraction in the Web page. It converts to an INM model tree with rules of extraction followed the INM data structure.

The features above should be converted to a rule template and inserted into INM model tree, so it can form an INM model tree with extraction rules. In this paper, we only establish extraction pattern to object class and role relationship class. The corresponding instance information can be extracted with the node of extraction rules attached to INM model tree.

1. The extraction of object class (root node)

For the root node of pattern tree, we set the URL as the DOM feature. For example, when we will extract instance information from Wuhan University, Wuhan University is the root noted of INM model tree. So the DOM feature is the website “www.whu.edu.cn”, semantic feature is the special word “university”.

**The extraction rules of root node instance:**

```plaintext
Create InstanceRule I_ University Name
{  
  DOMRule Equal (The homepage URL);  
  SemanticRule Equal ("University Name");  
}
```

2. The extraction of object class(not root node)

For the object class which is not root node, the DOM features are designed as Xpath and separate label, visual features are the position and font, semantic feature is the special word “school” or “department”.

**Create** InstanceRule I_Object class

```plaintext
{  
  DOMRule1 Equal (Xpath );  
  DOMRule2 Equal (separate label)  
  VisualRule1 Equal (positionoff(x,y))  
  VisualRule2 Equal (font)  
  SemanticRule Equal ("school" or “department”);  
}
```

3. The extraction of role relationship class

For the role relationship, the DOM features are designed as Xpath and separate label, visual features are the position and font.

**Create** InstanceRule I_Object class

```plaintext
{  
  DOMRule1 Equal (Xpath );  
  DOMRule2 Equal (separate label)  
  VisualRule1 Equal (positionoff(x,y))  
  VisualRule2 Equal (font)  
}
```

Insert rule is used to Create rules defined and inserted into the corresponding model of tree nodes, that is the mapping process. By performing the Insert rule, the pattern tree became a model tree with rules. And here is the illustration of the specific writing.
<table>
<thead>
<tr>
<th>University Sites</th>
<th>Number of colleges</th>
<th>Number of department</th>
<th>Number of faculty</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.whu.edu.cn">www.whu.edu.cn</a></td>
<td>40</td>
<td>286</td>
<td>3627</td>
</tr>
<tr>
<td><a href="http://www.pku.edu.cn">www.pku.edu.cn</a></td>
<td>64</td>
<td>375</td>
<td>3812</td>
</tr>
<tr>
<td><a href="http://www.hust.edu.cn">www.hust.edu.cn</a></td>
<td>47</td>
<td>302</td>
<td>3284</td>
</tr>
<tr>
<td><a href="http://www.shufe.edu.cn">www.shufe.edu.cn</a></td>
<td>22</td>
<td>158</td>
<td>1010</td>
</tr>
<tr>
<td><a href="http://www.urbe.edu.cn">www.urbe.edu.cn</a></td>
<td>23</td>
<td>149</td>
<td>932</td>
</tr>
<tr>
<td><a href="http://www.sdu.edu.cn">www.sdu.edu.cn</a></td>
<td>43</td>
<td>324</td>
<td>2147</td>
</tr>
<tr>
<td><a href="http://www.whut.edu.cn">www.whut.edu.cn</a></td>
<td>25</td>
<td>166</td>
<td>705</td>
</tr>
<tr>
<td><a href="http://www.znufe.edu.cn">www.znufe.edu.cn</a></td>
<td>20</td>
<td>132</td>
<td>655</td>
</tr>
<tr>
<td><a href="http://www.shufe.edu.cn">www.shufe.edu.cn</a></td>
<td>14</td>
<td>96</td>
<td>629</td>
</tr>
<tr>
<td><a href="http://www.xmu.edu.cn">www.xmu.edu.cn</a></td>
<td>57</td>
<td>345</td>
<td>2587</td>
</tr>
<tr>
<td><a href="http://www.nwu.edu.cn">www.nwu.edu.cn</a></td>
<td>23</td>
<td>141</td>
<td>1958</td>
</tr>
<tr>
<td><a href="http://www.nju.edu.cn">www.nju.edu.cn</a></td>
<td>36</td>
<td>233</td>
<td>1152</td>
</tr>
<tr>
<td><a href="http://www.tju.edu.cn">www.tju.edu.cn</a></td>
<td>24</td>
<td>157</td>
<td>845</td>
</tr>
<tr>
<td><a href="http://www.swufe.edu.cn">www.swufe.edu.cn</a></td>
<td>21</td>
<td>63</td>
<td>1231</td>
</tr>
<tr>
<td><a href="http://www.habu.edu.cn">www.habu.edu.cn</a></td>
<td>26</td>
<td>182</td>
<td>1032</td>
</tr>
</tbody>
</table>

1) Semantics: insert the rules of the root to the root node
   InsertStatement = INSERT INTO CLASSNODE rootName
   INSTANCERULE instance RuleName;

2) Semantics: insert the rules of Object class to Object class node
   InsertStatement = INSERT INTO OBJECT CLASSNODE Object class RULENAME Object class rule;

3) Semantics: insert the rules of Role relationship class to Role relationship class node;
   InsertStatement = INSERT INTO ROLE RELATIONSHIPCLASSNODE Role relationship class RULENAME Role relationship class rules;

When the college record set web page is the same one of department record set, it needs to confirm the relation between the college and department in HTML web page. As shown in Figure 5, a collection of the departments under the title of schools. So it can match the couple of schools and corresponding departments.

![Figure 5. College and department record set in the same web page](image)

V. EXPERIMENTS

In this section, it has carried out the corresponding experiments according to the above research methods. It includes analysis of data extraction of colleges, departments and faculty. The experiments were performed on a 2.66GHz PC machine with 1GB main memory, running on Windows XP. We collected 15 university web sites, and the following table shows the distribution of object class and role relationship class entities in these universities.

This paper use Precision, Recall and F-measure to evaluate experimental results.

\[
F\text{-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}
\]

The entities extraction of college and department, as well as faculty are started from a university seed URL, so this experiment needs to be divided two steps to analysis the effort. First, to find the web pages which contain entities information, and next to extract data from the obtained web pages.

<table>
<thead>
<tr>
<th>Method of this paper</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td>96.5%</td>
<td>97.3%</td>
<td>96.9%</td>
</tr>
<tr>
<td>Department</td>
<td>93.6%</td>
<td>94.8%</td>
<td>94.2%</td>
</tr>
<tr>
<td>Faculty</td>
<td>82.1%</td>
<td>79.5%</td>
<td>80.8%</td>
</tr>
</tbody>
</table>

On the aspect of Focused crawled, this paper method has been compared with Best first search [21]. Table 3 told that the result of this paper method mentioned is generally better than Best first search. The method of Best first search obtained results using calculate Relevancy of Topic keyword and the Web page crawled, while this paper add the visual layout features and URL anchor keywords to enhance the recall rate and precision rate.

<table>
<thead>
<tr>
<th>Method of this paper</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td>89.4%</td>
<td>72.6%</td>
<td>80.1%</td>
</tr>
<tr>
<td>Department</td>
<td>75.1%</td>
<td>70.3%</td>
<td>72.6%</td>
</tr>
<tr>
<td>Faculty</td>
<td>67.3%</td>
<td>70.8%</td>
<td>69.1%</td>
</tr>
</tbody>
</table>

Table 4 Indicates that the results of colleges and departments are very good, the F-measures are 97.7% and
93.3%, and the results of faculty are not better than object class. In the extraction process of the record set, object extraction results are better than role relationship because colleges and departments in the webpage of the arrangement regularly, the selected features are enough for extraction. But in role relationship extraction, the names of faculty are too much so here can’t design semantic feature, and it will affect the extraction results. Overall, the method can achieve the university webpage data extraction for INM.

VI. SUMMARY

This paper has proposed an entities extraction method for Semi-Structured database. According to the different forms of object class and role relationship class data in the web pages, this paper uses the visual features, DOM features and semantic features to develop extraction template. Then the feature templates are inserted into the structure of INM deserialization to help for extracting data. Finally the results of the experiments are pretty good. But unfortunately, there is some degree of manual intervention during the data extraction. And it has an effect on extraction efficiency. Next step is the study for automatic data extraction method based on the INM database. In addition, this paper tells nothing about extraction of attribute, and this is also the next stage of the research.

REFERENCES