Automatic Extraction of Semi-structured Web Data

Fang Dong[1], Mengchi Liu[1] and Yifeng Li[2]

[1]*State Key Lab of Software Engineering, School of computer, Wuhan University, Wuhan, China*
[2]*School of computer, Carleton University, Ottawa, Ontario, Canada*
[1]*whdongfang@163.com,* [2]*mengchi@scs.carleton.ca,*[3]*liyifeng666@126.com*

Abstract

As a huge data source the internet contains a large number of valuable information, and the data of information is usually in the form of semi-structured in HTML web pages. In order to extract the web data and organize the data with the relationships which are similar to the real world, this paper has proposed a method for automatic data extraction from the web. With the combination of keywords and database content matching, the target web pages which contain valuable data will be crawled. Via HTML structure and visual features, extracting the data from the web pages crawled. Eventually, the data been extracted will be integrated to the structure of information network model. Experimental results indicate that this method can be able to apply to semi-structured data extraction in the web, and this paper has provided positive significance to extraction and manage semi- structured web data.

Keywords: *Automatic Extraction, Semi-structured Data, Web Data, Information Network Model*

1. Introduction

With the explosive development of World Wide Web, more and more information can be extracted from the huge data source of internet. The web data like the objects in the real world is also connected by multiple relationships. But interrelated web data is always in different web pages, and even in the same web page the data lies in the HTML page with semi-structure. Since the relational database is widely used during a long time, the most data extraction previous orients the relationship model. But it is increasingly difficult for traditional database model to manage the web data with complex relationships. If the web information can be extracted and integrated by relationships that are closed to the real world, it will be helpful for managing the valuable data. Figure 1 is the typical structure of a university. As can be seen in the figure, a university contains teaching unit, management, research department. In the teaching units, there are schools/colleges, laboratories and departments, meanwhile schools/colleges also contain departments and laboratories. There are a lot of people as different roles in these organizations, such as presidents, professors, lecturers and students. Therefore the university has various linkages between different objects, while relational database storing data in the form of bivariate table has much difficult to express these objects with complex and diverse relationships. Corresponding to the real world objects and relationships, the data exists in semi-structured form in different web pages of internet. How to automatically extract the data scattered in different HTML pages and organize it according to the structure of reality is a major challenge for data extraction. This paper has proposed a method to automatically extract semi-structured web data in this background, and integrate the data to a database which can express context and complex relationships.

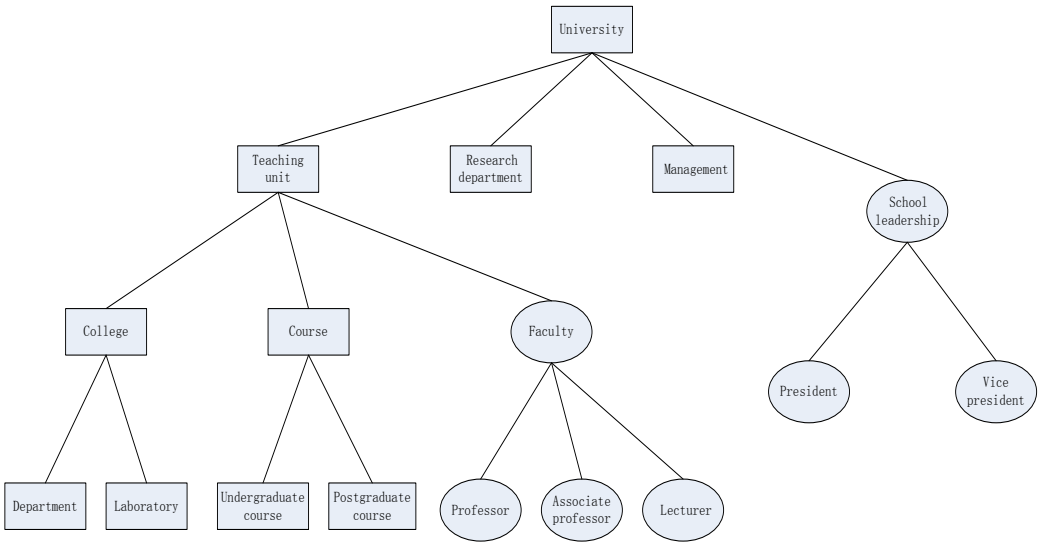


Figure 1. The Structure of the Real University

In recent years, with the rapid development of the internet technology, web becomes the data source with abundant information. The data extraction from the web pages efficiently according to different application background is the research target of web data extraction. The task is divided into two parts: one is focused crawler, and the other is the data extraction after finding the target web pages. Focused crawler [15, 16] is a program that automatically gathers web pages by exploring the linked structure of the World Wide Web. There are many approaches of focused crawler mentioned in the literatures [18-20]. Web data extraction has been introduced in the literature [2]. Currently there are a lot of data extractions for structured database. For example, according to the data extraction of different objects, the literature [3] talks about entity extraction, the literatures [4-6] are entity relation extraction methods, and the literatures [6, 7] are extraction for event. But most of these extraction results will integrate into two-dimensional table for relational database. There are many methods for web data extraction, such as data extraction based on natural language analysis, the typical systems are SRV [8], WHSH [9] and KnowItAll [10]. A web text will be divided into a plurality of sentence, and each sentence will be marked according to the grammar. Then to extract data using grammatical structure of the sentence and semantic information matching rules. The extraction method based on the analysis of natural language is suitable for the web document containing a complete sentence and clear grammatical case, so the extraction rules express finitely and it is difficult to extract the complex objects. In the current research of data extraction, there are wrapper induction tools [13, 14] which generate extraction rules derived from some given training examples. These tools are more suitable for HTML document because they rely on formatting features of the data in the web pages. The literatures [1, 11, 12] also mention the analysis methods of data extraction using HTML tags structure characteristics. Because the HTML web page labels contain some construction information, which is quite advantageous for data extraction in the web pages neatly. This paper also has used this method, but in view of the large number of university sites and diverse characteristics of a single HTML page, extracting data only with this method will be difficult, so here content matching will be combined to finish the work.

2. Problem Statement

Figure 1 shows the information of university in the real world, including institutions and people. Figure 2 contains some web pages of Carnegie Mellon University corresponding to the university in Figure 1. The task is to extract important data of Figure 1 from the data source of web corresponding Figure 2 and to organize data with relationships of reality. In the university, the most important organization is teaching unit including the college, department and laboratory, the most important people are the faculty members and they work in different teaching unit with different identities.

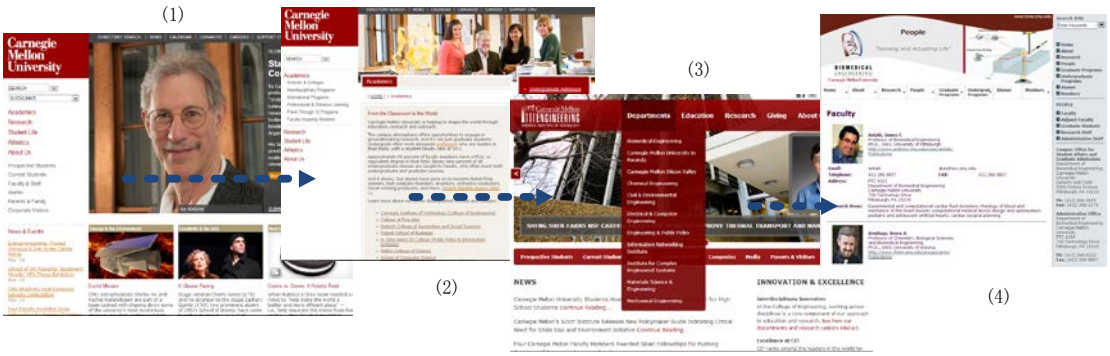


Figure 2. Instance of Carnegie Mellon University

Webpage (1) is the homepage of Carnegie Mellon University, and by keywords of “Academics” it can reach the webpage (2) which contains schools/colleges record set. All hyperlinks of schools/colleges are arranged neatly in the webpage and via these hyperlinks each college homepage can be reached. Webpage (3) is the homepage of Carnegie Institute of Technology (College of Engineering). In this web page there is a text tag “Departments” guides the menu of departments, which can point to the homepages of all department in this college. If entering “Biomedical Engineering” department homepage, the text hyperlink “faculty” guides to the web page of faculty record set, that is webpage(4) which exhibits information of faculty who serves in this department.

Through the typical examples as we can see, the extraction of this paper is mainly divided into two parts：

（1）Crawling target web pages from the university home page.

（2）Extracting colleges, departments and faculty from crawled web pages.

According to the characteristics of web data in the web pages, we have proposed the system of WEB-UIE for automatic extraction of semi-structured web data. As the university hyperlinks usually have obvious keywords, such as "Schools" and "Colleges", "Departments", "Faculty & Staff". Using visual features to identify relevant web pages, and then to confirm the subject web pages containing the data through following analysis of HTML pages. After obtaining the web page of data collection, extracting data with visual features, combined with the structure characteristics of HTML and the content of statistics together according to the institutes (College and department) and people (faculty) separately. For the college and department of extraction, in addition to the names of the college and department extracted, we must obtain the links pointing to the homepages, so that it will help the next step of the extraction work.

3. Our Approach for Data Extraction

3.1. Focused Crawling with Keywords Inspiration

To extract the information of teaching units and faculty, the first step is to find out the HTML web pages containing the data, then to analyze web pages to extract data based on the results.

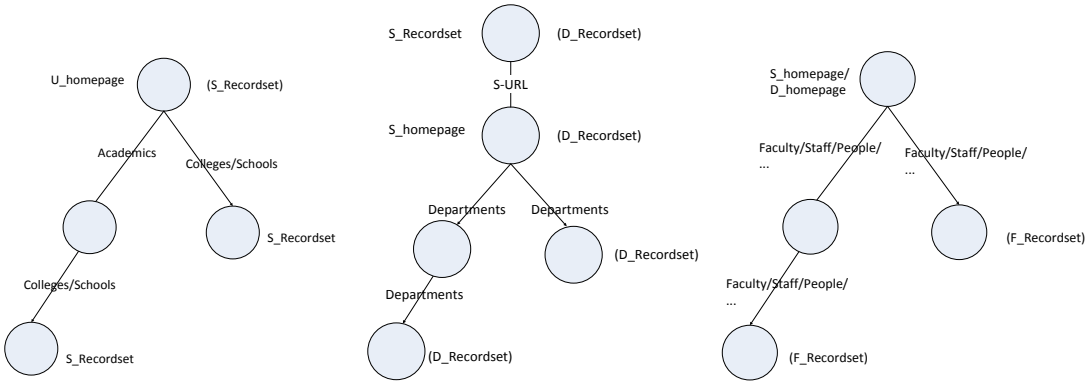


Figure 3. Keywords of Text Links

Figure 3 shows the keywords of text links between university homepage (U_homepage) and college data set web page (S_Recordset); college homepage (S_homepage) and department data set web page（D_Recordset）; college homepage (S_homepage) and faculty data set web page (F_Recordset); department homepage (D_homepage) and faculty data set web page (F_Recordset). As the keywords of text links in the university are quite obvious, the experiment result is good and the F-measure in the college and department reaches a value of 97.9% and 96% respectively. But the keywords of text links to faculty data set web page are more, such as “People & Faculty”,“Faculty Directory”. So the result of faculty data set web page is not better than that of college and department, but still can reach 82.3%. If it can’t find out these mentioned keywords, to confirm the target web pages directly through data area location and analysis of the web page structure and content.

3.2. Data Region Location

After collecting candidate target web pages with keywords, the next step is to determine which web pages really contain data, which belong to the noise web page. In HTML web pages of university, the structures of the web pages which contain information of teaching units and faculty are quite different. This paper designed a data area location method for college (or department) and faculty. If the data area can be located, the web page will be considered as the target web page; on the contrary the web page will be considered as the noise web page. In the past data location during the data extraction, generally the data to be extracted will occupy large area and in the center of the web page. It can help us to locate the data. But in this paper, as shown in Figure 2, the college (department) data and faculty data are not in the center of the web page and they do not occupy the largest area, even the data region is not continuous. So to locate the data is more complicated here.

3.2.1. Teaching Units: The traditional methods for data regions location is identifying the whole data region and then obtaining the precision region through boundary denoising,

because the record usually occupy large area and in the middle of the web page. But the teaching unit data does not meet the above two conditions, so it is impossible to locate data region though visual inspiration. Fortunately, these data are usually arranged tidily in the HTML web page and it brings about beneficial help to our task. In web pages containing colleges and departments, the data to be extracted can be distributed anywhere in the web page, but in clear format. So to collect data set following the definition 1, then to determine which data set is need to be extracted through database matching in definition 2.

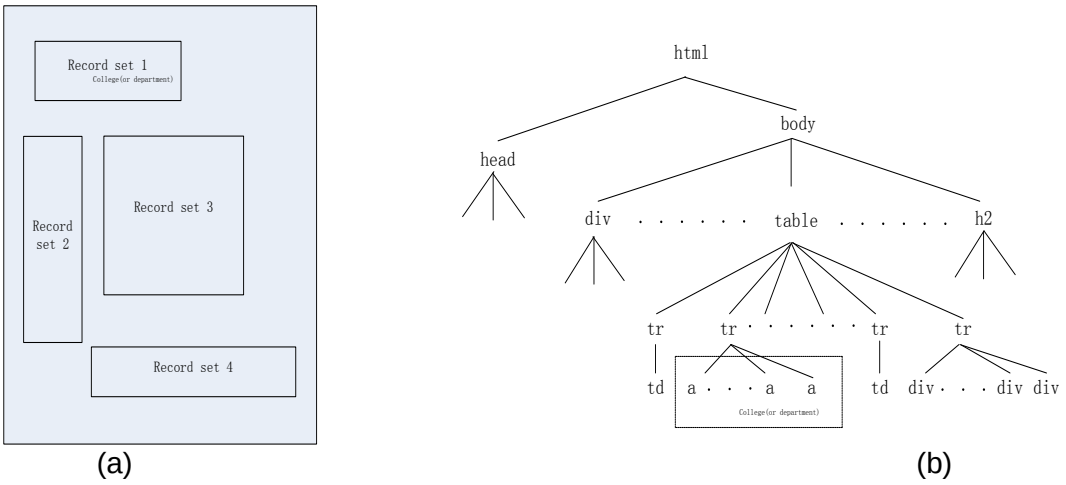


Figure 4. The College (or department) Data Record Layout

Definition 1 Record Set.

The HTML tags which encapsulate 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 = {h1 ,…,hs}.

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.

In set of H = {h1 , …, hs }, when s>4,it indicates that this set of data may be the collection of college or department data items. The reason is that the university (or college) usually contains 4 colleges (departments) at least. Obviously the data item is less likely to be college or department name when items of data set are less than 4, but if the data items are more than 4, it cannot prove that the data item is more likely to be college or department name.

Definition 2 Database Matching.

Setting up college and department related words database colleges.data, departments.data. The data in the database is represented as C={ c1,c2,⋯,cn},D={ d1,d2,⋯,dn }in which ci and di are college name and department name, h r is a set containing the string h r={h r1,…, h rt.}. Matching data items in the colleges.data (or departments.data) with the character in set h1…hs when s>4 in definition 1. Let k=0, if $\forall h_{ri} \in C\ (1 \le i \le t)$ or $\forall h_{ri} \in D\ (1 \le i \le t)$, that indicates hr is college record or department record, then set k+1 in H. When k/s>0.65, think all the elements in H are college record or department record.

Figure 4 has explained the matching mechanism of school record set and school name

database-colleges.data. Colleges.data contains 568 college relative words, here need to pay attention to determine the college according to the semantics when establishing database. For example, computer science cannot be divided into two words because the whole phrase is a college name. Meanwhile, there is a college named science which is different from the college named computer science. In the same way, there is a department database will be set for matching the department names-department data. Thus the college and department data area can be located.

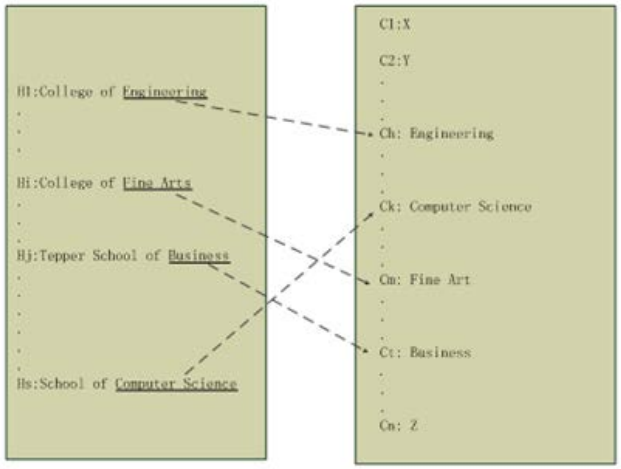


Figure 5. Database Matching of College

3.2.2. Faculty: Faculty is the most important people in the teaching unit of university. In this paper, the faculty extraction data includes instance name of faculty and the title. So we need to get the list of faculty name and corresponding title.

Definition 3 Repeating Units.

Since the data is usually at relatively neat way to show in the HTML pages that contains faculty records. The data is also encapsulated in the same structure of HTML labels. Therefore, more than one of faculty records in an HTML page will appear multiple repetitive structures of tag sequence groups. Repeating structure with the same tag sequence group is a repeating unit, there is a sequence group k repeating units $R = \{ri \mid i = 1,2, ..., k\}$.

Figure 6 has shown in a HTML page all the children nodes in a parent node. There are 11 children nodes under <tbody> nodes as a parent node. The sequence of its child nodes can be represented as in Figure 6.It indicates the tag sequence of the first level child in a parent node. The faculty records are usually in a similar display in HTML format, so two people records corresponding sub-tree structure are more similar than others. We have used similarity algorithm tree mentioned in the literature [1]. Here we set the similarity threshold as 0.8, a large number of experiments show that most people sub-tree can be distinguished with this threshold, and if Similarity >0.8, two sub tree will be considered as similar structure.

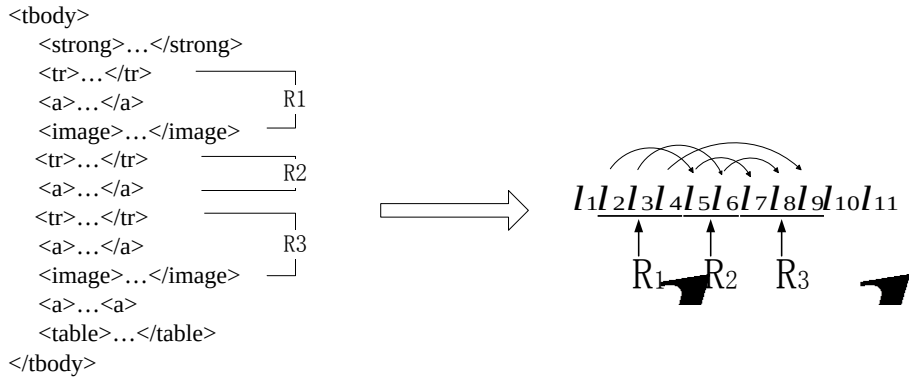


Figure 6. HTML Web Page Labels and Repeating Units

Tag sequence $l= l_1\ l_2\ l_3\ l_4\ l_5\ l_6\ l_7\ l_8\ l_9\ l_{10}l_{11}$ represents the sequence of 11 labels under a parent node, and it is divided into three repeating units r1, r2 and r3, each repeating unit is a candidate people records. Starting from l_1 label, the label with the following sequence similarity is calculated. If no similar sub trees, then Starting from l_2 label to re-calculated with the following sub trees and l_3、l_4 are found not similar to l_2.But l_5 Similarity (l2, l5)> 0.8, then this paper confirm they are two similar sub trees. Meanwhile $l_2l_3l_4$ is identified as a possible repeating unit. l_2 will continue to be calculated with the following sub trees, the fact that l_7 is similar to l_2 illustrates $l_5\ l_6$ will be temporarily divided into a repeating unit r2.The repeating unit r1 and r2 are attached to a parent node to calculate the sub trees similarity, the result has determined r1 and r2 are repeating units, while recording the start and end positions of r1. The repeating unit r2 will be calculated with all sub trees and be found no similar sub trees. Taking all labels sequence after l_7 as a sub tree to compute the similarity, if not similar, carving up all the remaining label sequence with the r1 length. The third labels sequence after l7 will be conformed a unit to be calculated with r1, the sequence is a candidate personnel records, denoted by r3. If repeating unit number k> 3 in a layer of a parent node, the result will be determined that these repeating units may contain people records. Algorithm 1 has explained the calculation of the repeating units.

Algorithm 1

Input: *ParentNode*//a parent node in HTML
Output: *RepeatN*// the number of repetitions
LoMark// the start and the end position of r
TChildren=getChildNodes (ParentNode)
RepeatN=0 ;i=0;
tempBegin = i;
1**for** *i* to number of sequence
2 *t=i*+1;
3 **while** *t*< number of sequence
4 and *Ti* is not similar to *Tt* **do**
5 *t=t*+1;
6 r_1=the sequence from*Ti* to*Tt*;
7 *Lunit*= the sequence length of r_1;
8 *LoMark*= the start & the end position of r
9 *tempBegin*=t;
10 **while** *t*< number of sequence
11 and *TtempBegin* is not similar to *Tt* **do**

```
12        t=t+1;
13        r2=the sequence from TtempBegin toTt;
14      if t< number of sequence
15      if r1 similar to r2
16      RepeatN= RepeatN+1;
17          else r2= the sequence from
18          TtempBegin to TtempBegin+Lunit;
19       LoMark= the start & the end position of r;
20            if r1 is similar to r2
21              RepeatN= RepeatN+1;
22            end if
23       end if
24    end if
25end for
26Returen RepeatN& LoMark
```

To establish a name database that is Name.data. Then extracting all text records encapsulated by the repeating unit with database matching which is mentioned in definition 3. Calculating the number of records containing name string in the database of Name.data, while marking the label and locations of names in the records. If the percentage of the repeating unit containing names records is greater than 0.65, it can identify this group of repeat unit is faculty data.

3.3. College and Department Extraction

After obtaining the data set of records, it is time to extract elements of H and corresponding hyperlink to the college or department home pages. Via database matching of definition 2, data records can be identified to be college (or department) data records or not. The elements h can be extracted from H set for college and department names.

For the college or department of the hyperlink extraction, it will be divided into two types:

(1) If the element h is separated by label <a>, the hyperlink corresponding to the h will be encapsulated in label <a>.

(2)If the element h is not separated by the label <a> but others, we need to find the nearest label <a> of college or department separate label in the parent or child nodes. In general the hyperlink of element h will be found through searching two layers.

3.4. Faculty Extraction

In the Chapter 3.2, after obtaining the repeat unit of faculty record, the names and titles of the records should be extracted. For person name, marking the separate labels in which data items match the name database. When the labels are the same with the obtained data labels and the Xpath of these labels in HTML also are similar, the data in these labels can be identified as faculty name. Sometimes the labels which contain the name will be also consists of other content. But the first letter of English names is capitalized and only a space separates two words of names. So it can help us to confirm the names of faculty. For the title, there are feature words-professor, associate professor, assistant professor, lecturer, Instructor corresponding to the name in each repeat unit, so using regular expressions to extract the relevant titles.

In the process of locating faculty data area, we think the number of repeat unit which is greater than 3 as faculty record. In fact the faculty record set displayed in the HTML web page not necessarily continuous, it will result in data loss. Therefore the labels with the same

Xpath with faculty record will be marked, calculating the similarity between two sub trees. If Similarity>0.8, it will be confirmed as person record.

4. Data Integration

As a kind of semi-structured database model, information network model can reflect the real world complex objects and relationships better than relational data model. This database can provide rich semantic information and it models according to the real world objects in a direct and natural way. The characteristics of the model decide the database can be better applied to the management of web data. Compared to the other databases, the INM database has great advantages in the management of university, it can represent various attributes of all the institutions and people in the university and the relationship between them. In the model the objects with similar properties are abstracted as class. In order to be able to represent the static properties and the dynamic properties, INM data will be classified as object classes and role classes. INM defines the relationship set RT={normal, contain, role-based, specific, identification, context, inverse}.

For explaining the management of university with database data, Figure 8 is the data structure of INM.

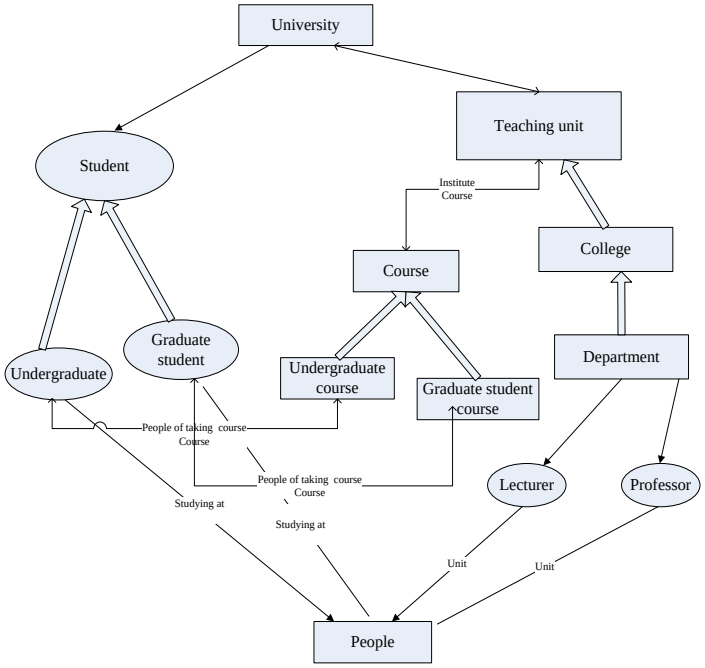


Figure 7. Data Structure of INM

As can be seen in the figure, a university usually contains multiple organizations and departments. Teaching unit is the most important sector of university and it contains most of the information. In the actual university web page, the hyperlink of this teaching unit will guide to web pages containing college and department data set. The university, teaching department, college and department belong to the object class, faculty belongs to the role relationship class. Following items are the relationships between them.

1. The university contains the teaching unit and the college contains the departments;
2. The teaching unit subsumes the college and department;

3. The relationship between faculty and college (or department) is in fact role played by the people, and that is the role relationship (role-relationship);

The extracted data is organized according to the INM data structure, the relationship to the real world can be obtained in this way. For example, Figure 8 represents insert statements:

Insert **University. College** Carnegie Mellon University. Mellon College of Science [
Faculty -> { **Professors:**{ Eric T. Ahrens, Bruce A. Armitage , David D. Hackney, Jeffrey O. Hollinger, Philip R. LeDuc },
Associate Professors:{ Ziv Bar-Joseph , Alison L. Barth, Peter B. Berget },
Assistant Professor: { N. Luisa Hiller, Sandra J. Kuhlman}
Lecturers：{ Daniel (DJ) Brasier ,}},
Contain **departments**:{ Department of Biological Sciences, Department of Chemistry, Department of Mathematical Sciences, Department of Physics}];

Figure8. Insert Statements of INM

5. Experiments

This section will discuss the effect of the method mentioned above through experiments. It includes the analysis of subject web pages crawler and data extraction.

5.1. Experimental Dataset

For college data extraction, this paper has collected 120 universities including 1687 colleges. In these colleges, choosing 120 colleges which are extracted exactly from the web as seeds URL, they will obtain 2259 departments. In the homepages of the obtained colleges and departments, 60 web pages were selected respectively out for seeds URL for faculty extraction, there are 3372 people finally.

5.2. Experimental Results and Analysis

In this paper, the traditional Precision, Recall and F-measure will be used to evaluate the experimental results. Since each college and department data in a university, as well as the faculty data in the schools and departments will be extracted from the homepage of the university, here are 2 steps for experimental analysis. For the first step, it will test the subject web pages that contain the valuable information, and the second step is to extract data from the crawled web pages. The dataset is considered to be a candidate extraction data only in the case of $s> 4$ in section 3.2.1 and $k> 3$ in section 3.2.2. Meanwhile the matching results of colleges data and departments. Data must be greater than 0.65, and the matching results of Name. Data must be greater than 0.75, in these situation the candidate dataset will be extracted. The threshold is set to 0.8 when calculate similarity between two DOM trees, and it will test the threshold of database matching and similarity in the followed section.

5.2.1. Analysis of subject web pages crawling: In the subject web pages crawling, the system WEB-UIE of this paper will be compared with best first search [17]. Input is the homepage URL of university and output is the identified web pages containing the valuable data. Through analysis of the data obtained in Table 1, WEB-UIE in college, faculty and faculty result are better than Best first search. Best first search method uses the similarity of keywords and crawled pages, and the extraction results will be quite good if the keywords are not many. The WEB-UIE in the faculty results are worst compared with college and department web pages, because this paper determine the scope of topic web pages based on

the keywords of URL, and the keywords of faculty are more cluttered than the college and department, thus affecting the results.

Table 1. Comparison of WEB-UIE and Best First Search Experiment

	WEB-UIE		
	Precision	Recall	F-measure
School	98.3%	97.5%	97.9%
Department	95.8%	96.2%	96.0%
Faculty	83.1%	81.5%	82.3%

	Best first search		
	Precision	Recall	F-measure
School	80.4%	82.6%	81.5%
Department	75.7%	73.3%	74.5%
Faculty	82.3%	80.8%	81.5%

5.2.2. Analysis of Data Extraction Results: In college, department and faculty information, the data which is need to be extracted will consists of college name, department name, corresponding homepage URL, people's name and his (or her) title. This paper will compare with method of MDR [1] in data extraction, the precision and the recall rate are based on the correct results of subject web pages crawled.

MDR extracts the data from HTML web pages according to the features of labels in the DOM tree. The experimental results have shown that for the school names, the faculty names and their titles WEB-UIE is better than MDR. Because to extract the data record only depending on the data label feature in DOM tree is not enough, WEB-UIE can determine the data to be extracted better using the method of database matching. But it needs to be improved in faculty extraction. Via analysis of faculty extraction error, we have found that the form of faculty data displayed in the HTML web pages is diversified. For example, records of faculty data in the same web page may not be in the same layer of a parent node, but in 3.3.3 chapter the extraction with repeating units only in the same layer sub trees. Meanwhile, in the subject web pages other record sets may be more than the record sets to be extracted. So in order to improve the recall rate we consider repeating unit more than 3 data blocks that may contain to extract data, but also expanding the noise information.

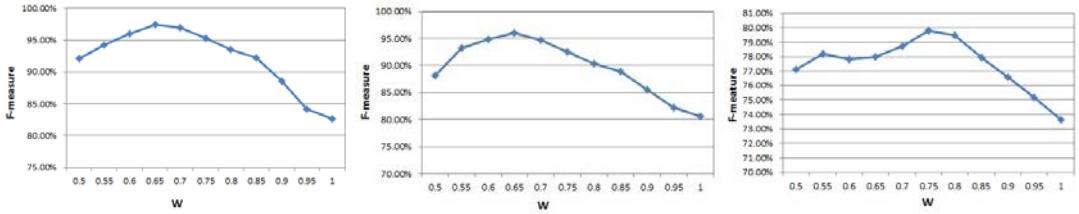
	WEB-UIE		
	Precision	Recall	F-measure
School name	97.3%	97.5%	97.4%
School URL	97.3%	97.1 %	97.2%
Department name	95.8%	96.2%	96.0%
Department URL	94.9%	95.6%	95.2%
Faculty	79.1%	80.5%	79.8%
Academic title	72.1%	70.5%	71.3%

	MDR		
	Precision	Recall	F-measure
School	85.4%	82.6%	84.0%
School URL	83.8%	82.6%	83.2%
Department name	85.7%	80.3%	83.0%
Department URL	83.9%	79.2%	81.5%
Faculty	70.3%	72.8%	71.5%
Academic title	69 .1%	68.5%	68.8%

Table 2. Comparison of WEB-UIE and DR Experiment

5.2.3. Effect of the Threshold for the Extraction: There are two thresholds to be tested, one is the ratio of database matching-W, another is the similarity of sub trees-S. After obtaining the record set, next step is to identify the ratio of valuable data among all repeating units through data matching. Here W is represented as the ratio, that is identifying the record set is to be extracted when the W reaches a value. During the faculty data extraction, the problem is that through calculate the similarity S to confirm two similar sub trees. Here the F-measure is the standard when experimenting with the colleges, the departments and faculty. Because our

approach is to extract the URL and people's title after obtaining the colleges, the departments and faculty name, the latter's F-measure lower than the former.



(a) College names (b) Department names (c) People's names

Figure 9. Effect of Threshold W to F-Measure

In the experiment of the threshold W, set similarity S as 0.8. As is shown in the figure 8, the W is in the range [0.5, 1]. During extracting the college and department data, W=0.65, the highest values of F-measure were separately 97.4% and 96%. But when extracting the faculty data, F-measure reached the highest value with W=0.75. During the experiment, if the threshold of W is too small, the noise information will be mistaken for valuable data to be extracted and the precision will be decreased; if the threshold of W is too big, it will lost a lot of correct information and the recall will be decreased.

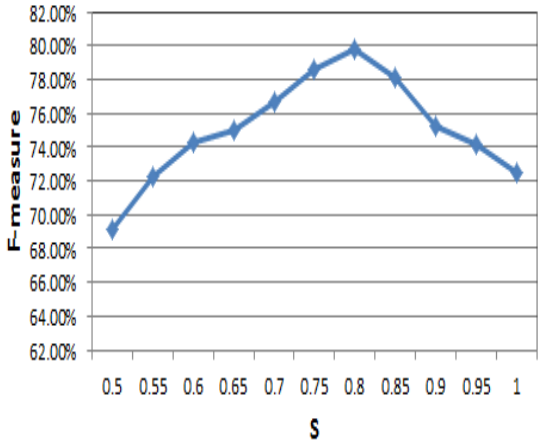


Figure 10. Effect of Threshold S to F-Measure

When testing the Similarity threshold S of sub trees, the W value is set to 0.75. As can be seen from the figure, F-measure will be highest when the tree similarity threshold is 0.8. Because in the web pages containing the faculty data, labels encapsulating person record typically have a similar format, then the sub trees will have high similarity.

6. Conclusions and Future Work

This paper presents a semi structured data extraction method based on network. According to different data objects and the object data in the webpage in the form of expression, we classify the extraction by object class and role relationship. Data extraction was performed using visual, database matching and HTML DOM tree structures. The experimental results show the effectiveness of the method. But some shortcomings and need improvements were

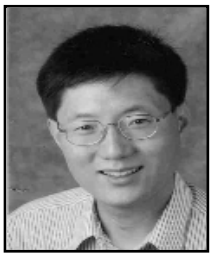
also found in the experimental process, for example, during the extraction process of teaching personnel records, because the data area personnel records in the HTML webpage cannot be determined, leading to the sub trees similarity of each layer of child nodes are calculated, so the efficiency needs to be improved. And, this is just the important institutions and personnel in the University and their mutual relations and extraction, the breadth and depth is not enough. It needs to expand to other institutions, faculty and extraction of the properties.

References

[1] B. Liu, R. L. Grossman and Y. Zhai, "Mining data records in Web pages", L. Getoor, T. Senator, P. Domingos, C. Faloutsos, Proc. Of the Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2003), Washington, ACM Press, **(2003)**, pp. 601-606.
[2] C. H. Chang, M. Kayed, M. R. Girgis and K. F. Shaalan, "A survey of Web information extraction systems", IEEE Trans. on Knowledge and Data Engineering, vol. 18, no. 10, **(2006)**, pp. 1411-1428.
[3] G. Weikum and M. Theobald, "From information to knowledge: Harvesting entities and relationships from Web source", Paredaens J, Van Gucht D, eds. Proc. of the 29th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems (PODS 2010).ACM Press, **(2010)**, pp. 65-76, [doi: 10.1145/1807085.1807097]
[4] R. Hoffmann, C. Zhang and D. Weld, "Learning 5000 relational extractor", J. Hajic, S. Carberry, S. Clark, Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010). ACL Press, **(2010)**, pp. 286-295.
[5] J. Zhu, Z. Q. Nie, X. J. Liu, B. Zhang and J. R. Wen, "StatSnowball: A statistical approach to extracting entity relationships", J. Quemada, G. León, Y. S. Maarek, W. Nejdl, Proc. of the 18th Int'l Conf. on World Wide Web (WWW 2009). ACM Press, **(2009)**, pp. 101-110.
[6] D. Ahn, "The stages of event extraction", Proc. of the Workshop on Annotating and Reasoning about Time and Events. Sydney, ACL Press, **(2006)**, pp. 1-8.
[7] S. Q. Li, P. Y. Liu, T. J. Zhao, Q. Lu and H. J. Li, "PKU_HIT: An event detection system based on instances expansion and rich syntactic features", Proc. of the 5th Int'l Workshop on Semantic Evaluation (ACL 2010), ACL Press, **(2010)**, pp. 304-307.
[8] D. Freitag, "Information Extraction from HTML: Application of a General Learning Approach", Proceedings of the 15th National Conference on Artificial Intelligence(AAAI 1998), **(1998)**.
[9] S. Solderland, "Learning Information Extraction Rules for Semi-structured and Free Text", Machine Learning, vol. 34, **(1999)**, pp. 233-272.
[10] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld and A. Yates, "Web-scale information extraction in KnowItAll (preliminaryresults)", Proceedings of the 13th World Wide Web Conference, **(2004)**, pp. 100-109.
[11] Y. Zhai and B. Liu, "Web data extraction based on partial tree alignment", A. Ellis, T. Hagino, Proc. of the Int'l Conf. on World Wide Web (WWW 2005). Chiba: ACM Press, **(2005)**, pp. 76-85.
[12] W. Jiying and F. Lochovsky, "Data extraction and label assignment for Web databases", Proc of the 12th Int Conf on World Wide Web, New York, ACM, **(2003)**, pp. 187-196
[13] N. Kushmerick, "Wrapper induction: Efficiency and expressiveness", Artificial Intelligence Journal, vol. 118, no. 1-2, **(2000)**, pp. 15-68.
[14] I. Muslea, S. Minton and C. Knoblock, "Hierarchical wrapper induction for semi-structured information sources", Autonomous Agents and Multi-Agent Systems, vol. 4, no. 1-2, **(2001)**, pp. 93-114.
[15] A. Patel and N. Schmid, "Application of structured document parsing to focused web crawling", Computer Standards & Interfaces, vol. 33, **(2011)**, pp. 325-331.
[16] A. Chandramouli, S. Gauch and J. Eno, "A Cooperative Approach to Web Crawler URL Ordering", Human Computer Systems Interaction, AISC 98, Part I, **(2012)**, pp. 343-357.
[17] F. Nczer, C. Pant and P. Srinivasan, "Topic-driven crawlers: machine learning issues [EB/OL]", **(2002)** May 15, http: / /dollar. biz.uiowa. edu/ ～fil/ papers.html.
[18] M. Kumar and R. Vig, "Design of CORE: context ontology rule enhanced focused web crawler", International Conference on Advances in Computing, Communication and Control (ICAC3"09), **(2009)**, pp. 494-497.
[19] W. Huang, L. Zhang, J. Zhang, M. Zhu, "Focused Crawling for Retrieving E-commerce Information Based on Learnable Ontology and Link Prediction", IEEC, International Symposium on Information Engineering and Electronic Commerce, **(2009)**, pp. 574-579.
[20] H. P. Luong, S. Gauch and Q. Wang, "Ontology-Based Focused Crawling", Information, Process, and Knowledge Management, 2009 (eKNOW '09), **(2009)** February 1-7, pp. 123-128.

Authors

Fang Dong received her M.Sc. in Computer Science (2010) from China University of Geosciences (Wuhan). Now she is a Ph.D candidate of Computer Department, Wuhan University. Her current research interests include Semi-structured database management.



Mengchi Liu received PhD in Computer Science (1992) from University of Calgary. Now he is a Professor and PhD supervisor. Since 2012, he is the director of State Key Laboratory of Software Engineering (Wuhan University). His main research interests include technology of database and intelligent information management system.

Yifeng Li received his Bachelor in Computer Science (2010) from Chongqing University. Now he is a postgraduate of Computer Science, Carleton University. His current research interests include Semi-structured database management.