2010 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010)
IPCSIT vol. 53 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V53.No.2.14

FpSp: A New Method for Web Extraction

XU Liping and SU Xiaohui[+]

School of Computer Science and Technology, Huazhong University of Science and Technology

Wuhan, China

Abstract—As the volatility of the web pages, the current Web extraction methods are very difficult to guarantee the robustness of extraction rules. This article presents a new statistics-based Web data extraction method, firstly utilizes the visual information in Web pages to classify pages to a narrow subject and a narrow information extraction region. We use the frequencies of data which occurs in the information block to extract data. It not only improves the robustness, but also improves the accuracy of Web extraction.

Keywords: automatic data extraction; information extraction; template generation

1. Introduction

Now the Web has become the oceans of information and sharing resources. How to extract data from the Web efficiently, to provide technical support for information retrieval and Web data mining ,has excited the widespread interest. Web information extraction (IE) is the method of extracting relevant data from HTML documents and storing the result as structured data. Wrapper is the first effective tool that used in Web information extraction. It can extract information which the users are interested in from the semi-structured HTML documents. However Wrapper requires the knowledge-based experts in the field, different procedures prepared for different site, and observation of the extraction rules to ensure the quality of extraction. Even so, the volatility of the pages and the complexity of the contents make it easy to Wrapper failure.

Current methods of automatic or semi-automatic Web information extraction are as follows, XWRAP [1] is a wrapper generator based on semi-automatic semantic HTML tag , it depends on the form of nesting and positioning, and it is suitable for the page containing significant regional structure but not for smaller Web site applications. WIEN and Stalker are the two systems doing better in WEB information extraction automatically. The common in WIEN and Stalker is to use machine learning techniques to extract rules, and the difference is that each system has a different rule extraction device. However, these methods above can not avoid manual intervention. This paper presents a new extraction algorithm FpSp (First partition Second pattern), we utilize visual information in Web pages to classify contents, and then collect statistical data of locations and frequencies in the block and store them in the template automatically. It improves the automation and accuracy of web information extraction, avoids the manual intervention almost.

2. Principles

Template is formed by the extraction rules, generally the rules can be got automatically by machine learning methods. Machine learning methods are generally built on the basis of training samples, by induction, learning and summarize to complete the rules identification of the web documents that whose rules has never been established. Usually, HTML page has some aggregation of information; similar information appears in

[+] Corresponding author.
E-mail address: xiaohuibeta@163.com

the HTML page as an informative block. The information which we need to extract often belongs to an information block. So information extraction can be divided into two steps. First, divide the whole page into a number of information blocks, then determine the block that contains the information to be extracted, it would avoid processing the entire HTML page as a basic object. Second find a suitable template in template library, utilize the data node path recorded in the template to combine a XQuery expression, meanwhile transform the information block to XML fragment, at last query XML fragment with the XQuery above. The principles of our extraction system shows in Figure 1:

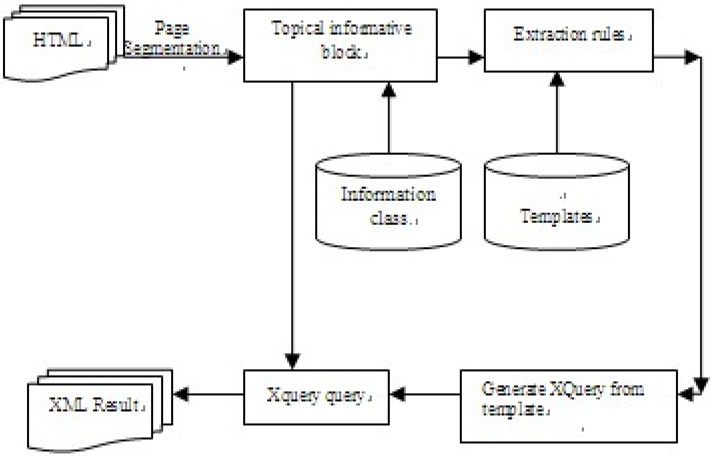

Fig.1 extraction system architecture

Extraction rules are generated in the sample training phase, it can re-used for structure-similar pages by simply testing whether the new web page is similar with the sample. In fact, because we have converted web pages to DOM tree, so it is enough to analyze the similarity of DOM tree.

3. Page segmentation algorithm

Web page authors generally organize content in a reasonable way, similar contents are placed together in an informative block. And a number of independent informative blocks make up the whole page as shown in Figure 2.The block location, size, links, pictures and other elements provide useful information to classify information and these features can be got through automatic analysis of program.

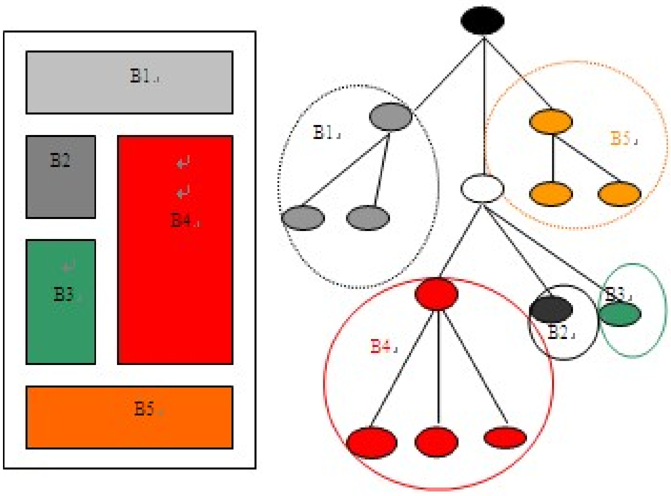

Fig.2. diagram of the page block

Suppose B={b1, b2, ..., bn} is an information block set, and bi represents the i-th block. And suppose all of the informative blocks are categorized into t classes, said with Ct. In the function S (i, m) = sim (bsi, bm), i = 1,2, ..., k, m = 1,2, ..., n; the first parameter bsi represents the s-th class and the i-th block, and the second parameter means th m-th block.

For Web page segmentation, the VIPS proposed by paper[9] is an effective segmentation algorithm, the VIPS segmentation algorithm makes full use of page layout features, such as font, color, size, etc.. The paper[10] also proposed a good page segmentation algorithm,it makes full use of the Web pages tags such an <table></table>.And the paper[11] gives some characteristic value of informative block such as **FontSize, FontWeight, InnerTextLength, InnerHtmlLength, ImgNum, ImgSize, LinkNum, LinkTextLength**,and all of these values form a feature vector v.In order to calculate the block similarity,we use the most basic calculation formula, that is, sim (v1, v2) = $\frac{v1 * v2}{|v1| * |v2|}$.Now we first provide the definition of block granularity.

Definition 1 block granularity It means the informative block which has a relatively difference from other in location, size, color, structure, etc. It is relatively independent in the semantics and takes together the similar data to a data region. Suppose a data nodes set W = (n_1, n_2, n_3, ..., n_k), call n_i in W a fundamental particle. For a partition on W = {W_1, W_2, W_3, ... W_n} , $\forall$ W_i, W_j belongs to W, there is $W_i \cap W_j = \varnothing$, and $\forall$ ni $\in W_k$, suppose V_{ni} is the feature items set of $n_{i,\ wk}$ is the feature items set of W_k, there is $V_{ni} \subset V_{wk}$.

We process block as the basic object, assuming that a page P can be expressed as a set of blocks, that is P ={b1, b2, ..., bn}. Then for all the samples S, with S = P1 ∪ P2 ∪ ... ∪ Pm.Cluster the information blocks with granularity and similarity, the S would be divided into t blocks of classes, all of the information class denoted as C = {C1, C2, ..., Ct}, which meets the following conditions:

$\forall$ bi $\in$ Cr, bj $\in$ Ct,record vector vi, vj feature vector for the vi,vj, then sim (vi, vj)> γ

The entire web page informative block classification process can be described as following pseudo-code:

```
Input: Web pages
Output: informative blocks of the web pages
//pNode means the current node that the process deals with
//nLevel mean the current level that the process deals with
Algorithm GenerateInfoBlock(pNode,nLevel)
 {
     if(Dividable(pNode, nLevel) == true){
       for each child of pNode {        //sub node
           GenerateInfoBlock (child, nLevel);
       }
   } else {
      if(CheckGranularity(pNode)==true) {
        Put the pNode into the block pool as a block;
        Put the feature vector of pNode into the vectorSet;
       }
   }
 }
Algorithm ClusterBlock()
{
  Queue q;
  for each block in block pool{
    mark blocki as Ci; //mark a type to each block in block pool
    q.enQueue(blocki);
   }
```

```
    while(true)  //cluster the blocks
    {
       maxsim= ɣ ;
        for any two feature vector vi, vj in vectorSet
       {
            if(maxsim< sim(vi, vj))
            {  maxsim= sim(vi, vj);
               k=i;
               u=j;
              }
         }
     if（maxsim= ɣ ）break;
    //merge two class blocks and return a new class
    c=MergeClass(blocki,blockj);
     q.remove(i);q.remove(j);
     q.enQueue(newblock);
     vectorSet.remove(i); vectorSet.remove(j);
     vectorSet.enQueue(feeatureVecotr of c);
    }
  }
```

1) training samples preparation: suppose W is the set of training pages, then after the segmentation all the blocks still meet the following conditions: B = (b1, b2, ..., bn), and b1 ∪ b2 ∪ ... ∪ bn = W.

2) page segmentation: we used the VIPS to segment the page,and the difference is we detected block segmentation to make sure if it is appropriate, and then calculated the characteristics vector of each informative block, put them together into two queues .

3) informative block categoried: the calculation process is constantly looking for two blocks with largest similarity, step by step,merging the two blocks, re-calculating feature vector of the new block.Here we assume the minimum similarity to merge the two blocks is γ, when γ value no longer changes,it indicates convergence of blocks classification for the informative blocks.And between the collection of information classification C = (C1, C2,, ..., Cm) and the set B = (b1, b2, ..., bn), there must be a mapping f : B → A, bi → Cj (i = 1,2, ..., m, j = 1,2, ..., m).

4. Template Definition and Generation

Definition 2: **extraction rule** Extraction rule is a <value,path> two-tuples, it is a mapping relation between the data to extract and the paths where the data locates in the web page. Value is a variable, path is some paths that meet the XPath[4] string expression. Xpath is a language accessing to XML documents, it can be used to traverse the elements and attributes of the XML documents.

Definition 3: **template** It composes of extraction rules R, data model M, and the command C composition. Extraction rules R is some two-tuples like <value,path>, the command C is a set C= {file, let, var, retval}, M is a data model that meets XML description specification.

The following example is a simple template:

```
<?xml version= "1.0" encoding= "UTF-8"?>
<template>
<file name='xmlname' path="C:\"/>
```

```
<let var='V1'  blockid="id"path="[path1][path2][path3]…"/>
<let var ='V2' blockid="id" path="[path1][path2][path3]…"/>
<let var ='V3' blockid="id" path="[path1][path2][path3]…"/>
<let var ='V4' blockid="id" path="[path1][path2][path3]…"/>
<retval>
   return <datamodel>
           <num> (data ($V1)) </num>
            <name> (data ($V2)) </name>
            <star> (data ($V3)) </star>
            <address> (data ($V4)) </address>
         </datamodel>
</retval>
</template>
```

The label <template> defines the entire template.The<file> labels indicates the result xml file name and path to store.The first value in tags <let> mean the variable value name of the extracted data, path is the xpath where the data locates in the xml fragment, path1 is the preferred path, path2,path3,etc are alternative xpaths.The <retval> labels provides the data model defined by users.

Given some information block B = {b1, b2, ..., bm}, B is the set of information blocks after clustering. First,we transform the information block to XML fragment and then put it into memory. Using XML's tree structure,the block bi is converted to a tree to calculate the similarity of the paths.

Definition 4: **path similarity** suppose p1 = $tag_0tag_1tag_2$ ... tag_m, p2 = $tag_0tag_1tag_2$ ... tag_n.The path similarity between two nodes is:

$$\mathrm{Sim}\ (p1, p2) = \frac{|\,\mathrm{same}(p1, p2)\,|}{\max(|\,p1\,|,|\,p2\,|)}$$

The function same (p1, p2) calculates the common part starting from the 0th tags, the function max (| p1 |, | p2 |) takes the longest path between path p1 and p2.

We design a record structure, used to store path of data ,the type of the path, the count of path and the probability of the path.

Record{Path; Type; Count; Probability}

As the template data model is given by the user, so the formation of the template is like filling blanks with the appropriate path, the generation process is as follows:

1) Randomiz a value i, make the data node xpath in bi set as the initial xpath set.

2) Traverse the informative block sample set S, transform each block in S to xml fragment for easy access. Traverse each node in the xml fragment by level, if the node is data node, compare the path of this node and every path in the xpath set, find one which has greatest similarity with the path of this node, if the similarity is 1, then count=count + 1,ff the similarity is greater than α and less than 1(α is a threshold value), then generates a new record, not changing the type, and setting count to 1, if the similarity is less than α, then generate a new record, settting type a new value and setting count to 1.

3) Repeat step 2 until all of the information blocks are processed.

4) Utilize the xpath information set gathered to cluster by type, then calculate the value of Probability by count of the type divided by the sum of all Count, then sort them by Probability. At last we determine the field of data model from the corpus to the type of data model represented by the word section, if the biggest Probability less than β, or less than the number appears, indicating that the type is too decentralized and not used.

5) Finally, fill the value of the xpath into the template according to Probability in a descending order using the correspondence of type and data item.

5. Experiment and Analysis

To measure the performance of an information extraction system, the recall R and precision P [8] are mainly based on.

$$\mathrm{R} = \frac{|\{\text{retrieved results}\} \bigcap \{\text{relevant results}\}|}{|\{\text{relevant results}\}|}$$

$$\mathrm{P} = \frac{|\{\text{retrieved results}\} \bigcap \{\text{relevant results}\}|}{|\{\text{retrieved results}\}|}$$

In general, the recall rate and precision interacts each other. For the same retrieving, the accuracy decreases as the recall rate increases, and vice versa. For comprehensive evaluating of system performance, people usually calculated the weighted geometric mean of recall rate and precision rate, that is F index, which is calculated as follows:

$$\mathrm{F} = \frac{(\beta^2 + 1)PR}{\beta\mathrm{P} + \mathrm{R}}$$

where, β is relative weight of the recall rate and precision. When β = 1, the two are equally important; β> 1, the precision is more important than the other; β <1, the recall is more important than the other. β values are generally 0.5,1,2. This experiment selects component informations as test cases, the results shows in table 1 and table 2. The experiment is carried out in two steps, first step is the generating templates,second step is the extraction of information based on the template genetated in first step.

TABLE 1 TEMPLATE GENERATION, XPATH PATH TAG ACCURACY

source of pages	samples	tag ccuracy
www.sstc.org.cn	115	78%
www.skycn.com	236	82%
www.onlinedown.com	168	77%
www.componentsource.com	127	74%
www.flashline.com	183	81%
www.51component.com	192	83%

TABLE 2 INFORMATION EXTRACTION STAGE, THE ACCURACY OF DATA EXTRACTION

source of pages	samples	R	P
www.sstc.org.cn	115	76%	82.3%
www.skycn.com	236	79 %	82.4%
www.onlinedown.com	168	78%	81.4%
www.componentsource.com	127	77%	79.5%
www.flashline.com	183	74%	69.9%
www.51component.com	192	81%	83.9%

The core technology of information extraction is how to discover the extraction rules. Most information extraction technologies can discover extraction rules from the training samples. This article first utilizes the visual information to segment web page, the make a statistic of frequencies of the paths data nodes in a informative block to extract information accurately, the whole process try its best to avoid manual intervention. However the template definition and generation algorithms l needs to be optimized, this will be further work in this article.

6. References

[1] Wei Han, David Butter, Calton Pu,“Wrapping Web data into XML,” ACM SIGMOD Conference on Management of Data:New York, vol. 30, pp. 33－38,Setptember 2001

[2] I. Muslea, S. Minton, C. Knoblock,“A hierarchical approach to wrapper induction,” In Proc. of 3rd Intern. Conf.on Autonomous Agents: Washington pp.190 -197, 1999.

[3] N. Kushmerick, D. Weld, and R. Doorenbos,“Wrapper induction for information extraction,” In Proc. of IJCAI, 1997.

[4] XPath Language 2.0 http: // www.w3.org/TR/xpath20/

[5] XQuery 1.0 http://www.w3.org/TR/xquery/

[6] ZHI Zong-liang,CHEN Shao-Fei “Optimized Web Information Extraction based on XQuery,” Journal of Computer Applications. Vol. 28, no. 1, pp. 152-158. Jan. 2008

[7] YANG Shao-Hua,LIN Hai-Lue,Han Yan-Bo, “Automatic data extraction from template-generated web pages,” Journal of Software Vol.19, no.2, pp.209-223.Feb. 2008

[8] Gaizauskas Robert, Yorick Wilks,“Information Extraction: Beyond Document Ret- rieval,” Journal of Documentation, Volume 54,Number 1,pp. 70-105,1998

[9] Deng Cai,Shipeng Yu,Ji-Rong Wen,Wei-Ying Ma, “VIPS: a Vision-based Page Segmentation Algorithm Technical Report, ” MSR-TR-2003-79 Nov. 1, 2003

[10] Shian-Hua Lin,Jan-Ming Ho,“Discovering Informative Content Blocks from Web Documents,” International Conference on Knowledge Discovery and Data Mining: Edmonton, Alberta, Canada, pp.588-593,2002

[11] Ruihua Song,Haifeng Liu1,Ji-Rong Wen,Wei-Ying Ma “Learning Important Models for Web Page Blocks based on Layout and Content Analysis, ” CM SIGKDD Explorations Newsletter vol.6,pp.14-23, 2004