A Hybrid Method for Extracting Deep Web Information

Yuanpeng Zhang[1 a],Li Wang[1 b *],Kui Jiang[1 c], Danmin Qian[1 d], Jiancheng Dong[1 e]

[1]Dept. of medical informatics, Medical School, Nantong University, 19 Qixiu Road, Nantong 226001, Jiangsu Province, China

[a]maxbirdzhang@ntu.edu.cn,[b]wangli@ntu.edu.cn. [c]kuij@ntu.edu.cn, [d]sandy8793@ntu.edu.cn, [e]dongjc@ntu.edu.cn

Keywords: information extraction, clinic expert information, domain model, block importance model, SVM

Abstract. Some previous works show that more than 60% of the information available on the Web is located in Deep Web database. Such information cannot be directly indexed by search engines.In this paper, a hybrid method, which is composed of a domain model and a block importance model is proposed to extract information in Deep Web.The domain model is used for classifying and identifying whether a form is a WQI. The block importance model is used for filtering noisy information in response pages. These two models are both compared with a rule-based method. The experiment results indicate that the domain model yields a precision6.44% higher than that of the rule-based method, whereas the block importance model yields an F1 measure 10.5% higher thanthat of the XPath method.

1 Introduction

The explosive growth of the Internet has turned the Web into one of the most important sources of information containinga large number of available databases. Recent studies have found that more than 60% of the information available on the Web is located in Hidden-Web databases that are accessed by special HTML forms. This information is known as the Deep Web [1]. Theonly way to accessthese databases is through web search interfaces (WQIs). Figure 1 depicts this process. A WQI is be composed of many HTML elements such as *<form></form>*, *<input>* and*<select></select>*, ect., as well as labels and attributes.Two issues generally need to be dealt with when extracting information from theDeep Web. The first issue is to identifyWQIs among forms, and the second is to extract information from response pages. For the first problem, a simple rule-based method was proposed by J Cope *et al.* to determine whether a web page contains WQIs [2]. This method puts forward three rules.The first one isthat a WQI should contain HTML element *<form></form>*, the second one isthat a WQI should contain HTML element *<input type=text>*, and the third one isthat a WQIshould contains some attribute-words like *"search"*, *"query"*. Unfortunately, this method needs further improvements because it cannot distinguish search engines from WQIs. For the information extractionissue, it is critical to filter noisy information like navigation information, advertisement information and version information. Yan Fu *et al.* adopted a series of rules to distinguish informative contents blocks from noisy clutters, and generalized public XPath for the problem [3]. This method was tested in five different web sites and resulted in average integrality of 92% and average accuracy of 83.2%. However, this method has a basic precondition that the web pages should have a common layout.Therefore, this method does not have general applicability.

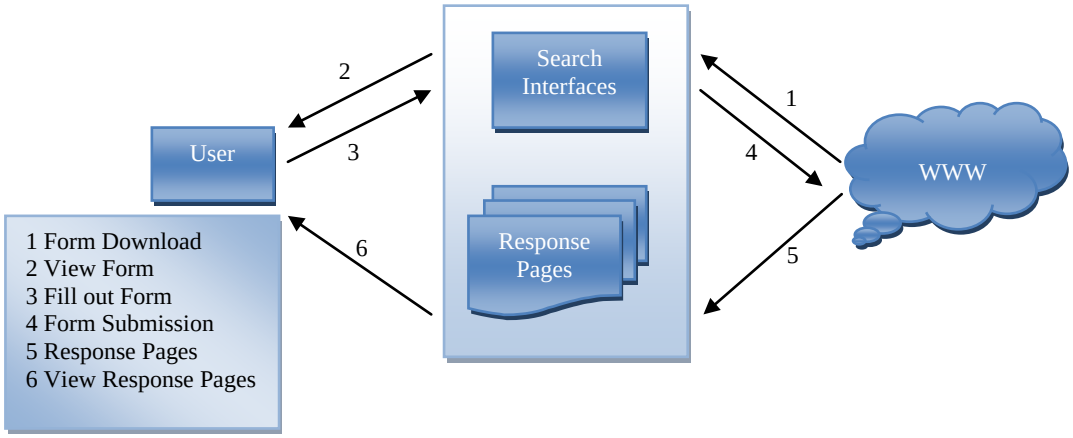

Fig.1. Process of deep web information extraction

In light of the above discussion, this paper proposes a hybrid method for extracting information from Deep Web databases. The main contributions of this paper are:

A domain model is defined to identify WQIs. The model can identify the domain to which an unknown interface belongs, and provide keywords to fill in WQIs.

A block importance model is proposed to filter noisy blocks in a Web page. Both content and spatial features are taken into account in this model.

2 Proposed method

On web pages, not all forms are WQIs; it is necessary to make judgments. We made some improvements to the rule-based method proposed by J Cope *et al*. Assuming that a form can be defined as a tuple that consists of five elements, i.e. $Form = \{\{C_1, C_2, C_3 \dots C_n\}, Action, Name, Method, URL\}$, where $\{C_1, C_2, C_3 \dots C_n\}$isa control list contained in $Form$,*Action*is anattributeof *Form*and its value is a URL that receivesdata from *Form*,*Name*is the name of *Form*,*Method* represents the submission method of *Form*and its value is *POST* or *GET*, and *URL* represents the location of *Form*, the following rules were derived.

- **Rule 1**. If*password box* belongs to$\{C_1, C_2, C_3 \dots C_n\}$, then *Form* is not a deep web search interface (*Form* may be a login form or a registration form).
- **Rule 2**. If *file upload box* belongs to$\{C_1, C_2, C_3 \dots C_n\}$, then *Form* is not a deep web search interface.
- **Rule 3**. If *textarea box* belongs to$\{C_1, C_2, C_3 \dots C_n\}$, then *Form* is not a deep web search interface.
- **Rule 4**. If the root directory of *Action* is different from that of *URL*, then *Form* is not a deep web search interface (*Form* may be a search engine form or ameta-engine form).

These rules can provide guidance onpreliminary screening of interfaces,but are not suitable for the judgment of some particular interfaces. Therefore, a domain model is proposedfor extractingtheir schema information.

2.1 Domain Model

Researchers from UIUC manually collected 477 WQIs in 8 areas through google engine and web directory service, and performed a statistical analysis of those WQIs and came to the following conclusions [4].

Attributes of each WQIs are finite.

Although there are numerous WQIs in each area, the vocabularies that depict the attributes of WQIs are convergent through clustering.

Based on these two unique features, a domain model that can describe the attributes of WQIs is proposed. The definition of the domain model is as follows.

Definition. The domain model is an ordered attribute tree that is defined as a 11-tuple, i.e. $DM = (V, v_0, E, \Delta, TP, N, Lb, Val, tf, R, \leq)$, where

V – a set containing all nodes

v_0–the root node, $v_0 \in V$

E – an edge set (linking parent node with child node)

Δ –a character set

TP – a function ($V \rightarrow \{(radio\ button, check\ box, text\ box, select\ list)^*\}$), which returns the node's type

N – a function ($V \rightarrow \{\Delta^*\}$), which returns the node's name

Lb– a function ($V \rightarrow \{\Delta^*\}$)

Val –a function ($V \rightarrow \{\Delta^*\}$), which returns the node's default value

tf– a function ($V \rightarrow \{N^*\}$) (N represents natural number), which returns the node's frequency of appearance in all search interfaces

R –a function ($V \rightarrow \{range,\ part,\ group,\ constraint\}$), which returns the relationship between node and its father node

$\leq$ –represents the order of nodes in the node set, $(u, v) \in \leq$ means that *u* appears before *v*

2.2 Construction of Domain Model

Based on the above definition, the construction of a domain model can be described as follows. Firstly, a WQI is chosen from one domain as an original domain model, then it is combined with other WQIs in this domain in order to enlarge and enrich the original one. This process is repeated until the domain model becomes stable. This combination process should comply with the following four rules.

Assuming u is a node belonging to the original domain mode and v is a new node, then

- **Add rule**: If the semantics of *v* is different from other nodes in the original domain model, then add a tree (*v* is the root node) to the original domain model.
- **Update rule**: If the semantics of *v* is similar to *u*, then update *TP list*, *N list*, *Lb list*, *Val list* of *u* with *TP*, *N*, *Lb*, *Val* of *v*.
- **Refine rule**: If the semantics of *v* is similar to *u*, and *v* contains some attributes that do not appear in *u*, then add *v* to the domain model as a child node of *u*.
- **Generalize rule**: If the semantics of *v* can generalize several nodes *{u1,u2…un}* in the original domain model, then add *v* to the domain model as a child node of these nodes' parent node; at the same time, take *{u1,u2…un}* as children of *v*.

For example, Figure 2 shows a typical example of two WQIs in the book domain, we can merge these two WQIs using the above rules. Figure 3 shows the merging result.

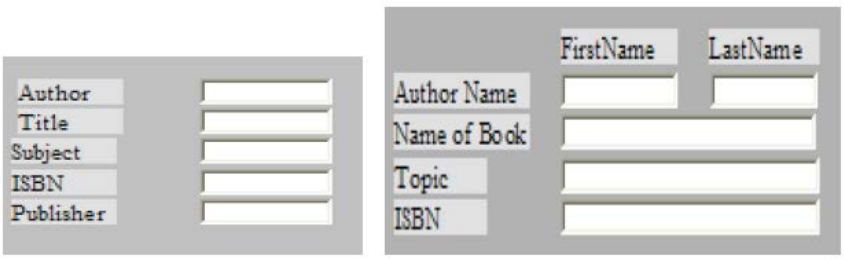

Fig.2. Two WQIs in book domain

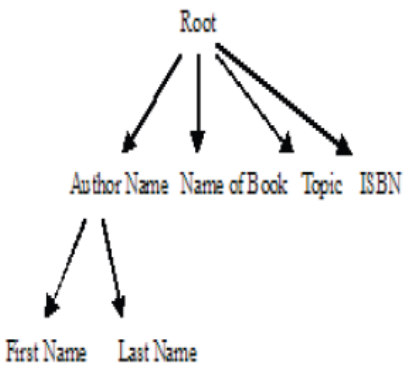

Fig.3. Hierarchical representation of two WQIs in book domain

2.3 Classification and judgment of search interfaces

Forms that cannot be judged by rule-based method should be classified and their attributes should be extracted based on the domain model.The process contains the following steps.

Suppose $\mathrm{Form} = \{\{C_1, C_2, C_3 \ldots C_n\}, A, N, M, U\}$ is a pending form, then extract attribute-words from$\{C_1, C_2, C_3 \ldots C_n\}$;

Standardize the extracted attribute-words, for example,remove stop words, restore word stem, filter illegal characters, etc.;

Calculate the similarity between Form and each domain model through Vector Space Model [5], Form belongs to the domain model with which it is better correlated;

Choose keywords from the domain model to which Form belongs, then fill in and submit the form. According to the returned webpage, a judgment can be made.

2.4 Information Extraction

The response web page often contains information that is irrelevant to the extracted theme. Such noise should be filtered to avoid topic drifting. A web page can be divided into different blocks such as navigation block, advertisement block, version block and content block,ect. Table 1 illustrates the importance levels of these blocks.

Table 1 Block importance levels

Level	Description
level1	Blockscontaining noisy information irrelevant to the main topic
level2	Blockscontaining relevant information but not about the main topic
level3	Blockscontaining information about the main topic

In order to deduce block importance from block features, two kinds of algorithms can be used. First one is rule-based algorithm, unfortunately, it is difficult to construct a rule function manually when faced with dozens of features. Second one is machine learning-based algorithm. Blocks are manually labeled as (x,y), where x is the feature of the block and y is the importance level of the block. Suppose T is the set of labeled blocks (training set), the learning process of this algorithm can be described as finding a function f(x) that minimizes$\sum_{(x,y)\in T} |f(x) - y|^2$.Clearly, the importance level (y) is a discrete value, so the learning process can be treated as a classification problem to be solved. Many available learning algorithms can be employed to tackle the problem. In this study, SVM [10] is chosen to construct the block importance model.

3 Experiment results and evaluation

3.1 Domain model experiment

In this experiment, we need WQIs and non-WQIs (non-searchable forms) to evaluate the performance of the domain model. One hundred and twelve WQIs are selected from 4 domain in TEL-

8. Table 2 lists the detail information of these WQIs. Non-WQIs are selected form http://www.searchengineguide and http://www.dmoz.org2. In addition, we also select some search engines and meta-engines as non-WQIs. Finally, we obtain totally 55 non-WQIs.

Table 2 WQIs from 4 domain

Domain	WQIs having default values	WQIs not having default values	WQIs	Percent
book	5	24	29	17.24%
movie	5	24	29	17.24%
automobile	9	9	18	50.00%

For comparison, a rule-based method is introduce and tested on the same data set. The accuracy of the domain model has been measured via 2 metrics: accuracy and precision. Suppose number of correctly identified WQIs are $\mathrm{right_{DP}}$, number of correctly identified non-WQIs are $\mathrm{right_{NDP}}$ and number of incorrectly identified WQIs are $\mathrm{wrong_{DP}}$. Let us now define the various accuracy metrics.

$$\mathrm{accuracy} = \frac{\mathrm{right_{DP}} + \mathrm{right_{NDP}}}{\mathrm{total}} \tag{1}$$

$$\mathrm{precision} = \frac{\mathrm{right_{DP}}}{\mathrm{right_{DP}} + \mathrm{wrong_{DP}}} \tag{2}$$

Table 3 shows the experiment result. The domain model achieved the best performance with Accuracy 98.81% and Precision 99.07%. The rule-based method performed worse than the domain model with Accuracy 81.44% and Precision 92.63%. Because the rule-based method is conservative, it can only achieve preliminary identification of WQIs. Interfaces like search engine forms, survey forms etc. are not easily identified by this method.

Table 3Result of deep web interfaces judgment

Methods	Interfaces	Number of correct judgment	Number of incorrect judgment	Accuracy	Precision
Domain model	WQIs	106	6	95.81%	99.07%
	Non-WQIs	54	1		
Rule-based method [4]	WQIs	88	24	81.44%	92.63%
	Non-WQIs	48	7		

3.2 Information extraction experiment

Two thousands and five hundred responsepages arecollected and divided into two groups, one group is using for training and another is for testing. Two learning method, L-SVM and RBF-SVM[6] areused forlearning algorithms. In addition, the method proposed by Yan Fu et al., i.e. XPatharetested on the same data set for comparison. To evaluate the performance of these methods, P, R and F1 measures aretaken using IR methods. Table 4 shows the experiment result.

Table 4Result of information extraction

Methods	Measures	Level1	Level2	Level3
XPath	P	0.645	0.701	0.842
	R	0.637	0.744	0.680
	F1	0.708		
Block importance model using L-SVM	P	0.753	0.788	0.830
	R	0.765	0.794	0.768
	F1	0.783		
Block importance model using RBF-SVM	P	0.775	0.802	0.891
	R	0.799	0.812	0.800
	F1	0.813		

RBF-SVM achievedthe best performance with F1 0.813. L-SVM achieved the performance with F1 0.775, worse than RBF-SVM. This result indicated that nonlinear features combination is better

than linear combination. Xpath achieved the performance with F1 0.708, worse than both RBF-SVM and L-SVM. The reasons accounting for this resultcan be described as follows:

First, when Web pages are large, it is practically impossible to ensure that all pages have the same layout. In addition, it is not easy to construct rule functions for the XPath method.

Second, SVM with RBF kernel has strong abilities of interpolation, it is adept in fetching local properties of samples.

4 Conclusions

A hybrid method is proposed in this paper for extracting Deep Web information. The method consists of a domain model and a block importance model. First, the domain model is used for judging whether a form is a WQIs.The experiment result shows that the model has a precision 6.44% higher than that of rule-based method. Then, the block importance model is adopted for noise filtering of a response page.The experiment result indicates that the model has an F1 measure 10.5% higher than that ofXPath method.

Acknowledgement

This work is supported by the National Science Foundation of China (No.81271668) and Jiangsu Natural Science Foundation of College and University (14KJB310014)

References

[1] M.K.Bergman,The deep web: surfacing hidden value, The Journal of Electronic Publishing 7 (2001), 3-21.

[2] J Cope,NCraswell, D Hawking. Automated Discovery of search Interfaces on the web. Proceedings of the l4th Australasian database conference, Adelaide, 2003, pp. 181-189

[3] Yan F, Dongqing Yang, Shiwei Tang. Using XPath to Discover Informative Content Blocks of Web Pages. Proceedings of the third International Conference on Semantics, Knowledge and Grid. SKG, 2007, pp. 450-453

[4] Fayzrakhmanov R, Information extraction from web pages based on their visual representation. Current Trends in Web Engineering, 7059(2012), 342-346.

[5] G.Salton, A.Wong, C.S.Yang, A Vector Space Model for Automatic indexing, Communications of the ACM 18 (2003), 613-620

[6] Burges C.A tutorial on support vector machinesfor pattern recognition.Data Mining and Knowledge Discovery,2(1998), 955~974.