*2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)*
***IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore***
*DOI: 10.7763/IPCSIT.2012.V51.69*

# Research Query Translation Based on Ontology Technology and Schema Matching

Gang Liu[+], Kai Liu and Yuan-yuan Dang

School Of Computer Science and Engineering Changchun University Of Technology, Changchun China

**Abstract.** Because of WDB's local interfaces exist the autonomy and heterogeneity,the query content of user from the integration of the query interface cannot be equivalently transformed into the local query, thereby reducing the accuracy of the query.In this paper, A domain ontology has been created based on the collection of each property of the local deep web query interfaces, followed by an analysis of the local query interface combined with the domain ontology to construction a general query interface which is able to meet the needs of users. When the user enter query conditions in the integration interface,the conditions will be matched with the matching table, if successful, go directly to the local query interface,failed then match the attributes and ontology,in order to simplify the query transformation process. schema matching was used in query translation module to examine the relationship of attributes between integration interface and local interface.

**Keywords:** deep web; query translation; interface integration; schema matching

## 1. Introduction

Web, according to the depth of information divided into Surface Web and Deep Web two parts, of which Deep Web refers to the Web information cannot be searched by a traditional search engine .the web informations in the deep web is stored in the database,to obtain the information must be queried by the query interface, and the results will be the form of dynamic page returned to the client.these dynamic pages do not exist the corresponding statically linked,so we cannot depended on the traditional search engine to get the informations in the deep web[1]. According to an Investigation in April 2004, Deep web contains approximately 450,000 pages database,and still continue to grow exponentially.

To effectively use the wealth of information in the Deep Web, more and more researchers began to focus the work of Deep Web data integration. The query translation is a core part of Deep Web data integration[2],which is mainly responsible for converting the query on the integrated interface to the relevant Web database query interface.

Since the autonomy and heterogeneity of WDB,making the local WDB entry form also has the corresponding autonomy and heterogeneity.Therefore the local query capacity for a given query was limited by the different of the query ertry form,resulting user's the query in unified query interface can not be transformed into local query interface equivalent. which causes the query result contains a lot of irrelevant information and repeated information.The purpose of query translation is to achieve maximum approximate query translation, to ensure precision in the premise of make the return results have characteristics with minimum redundancy that is to find the smallest superset of the desired results.

## 2. query translation frame design

Deep web query interface automatic filling framework shown in Fig. 1,first of all in order to establish the core domain ontology through the collection deep web of attributes of local database entry form,and in the form attribute extraction and scheme matching process to gradual improvement of the core ontology. Then create an common

[+] Kai Liu. Tel.: 18010156796.
*E-mail address:* liukai198612@gmail.com.

integrated query interface base on domain ontology and attributes of local deep web entry form,while ceate a property matching mapping to response the corresponding from the attributes of integrated query interface to deep web local entry form attributes . After the user submits a query, first check the legality, and then through the query translation processing into local query interface which the query can be adapted.Achieved by the transparent to the user query interface to all the local deep web query entry form the forward and rewritten.

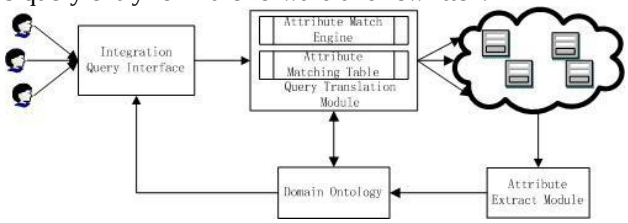


Fig. 1. query translation frame

# 3. Key Technologies

## 3.1. domain ontology

Ontology is a initial philosophical concept ,concerned about the abstract essential of objective reality .In the artificial intelligence domain, the first definition is given in Ontology is Neches and other authors[3],and them defines the Ontology as "An ontology defines the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary" .Because the process of building ontologies is different,generally believed that,Gruber in 1995's five rules has more influential: the clarity and objectivity, integrity, consistency, maximum one-way scalability, minimum binding.

In this paper,through the process of building a domain ontology,and extract the attributes from deep web form to study the relationship between words,to determine the division of related words, the concept of division including:synonyms relations, including relations , hyponymy relations and so on.

Domain ontology as a knowledge model to describe the concept of specific areas and the relationship between the concepts.In this paper,in order to improve query correctness and accuracy of conversion,we use of the collected query form attributes, to create domain ontology.

In order to reduce the participation of experts in the field,through this process by query translation,the ontology is constantly expanded and adjusted automatically. For example in attributes match mapping,the position of the attribute will be based on the number of query translation matching and accelerate the speed of query.

## 3.2. Schema Matching

The autonomy of the web makes the database even if the same or similar semantic attributes in different local query interfaces have different labels, different data formats and organizational structure. In order to be able to put user input query parameters correctly switch to the parameters of the local deep web query interfaces, this process of transformation in the query using a pattern matching.

Pattern matching refers to a given multiple mode, according to information that can be found between semantic equivalent model of mapping between members correctly process, it contains two matching type: simple matching and complex matching[4].Simple matching refers to the property of semantic matching between the 1:1 mapping. Such as the author and the author, the semantic match between the relationship is 1:1. Complex matching refers to attribute between the 1: M matching, or M: N matching, that is, a mode to another mode of M attributes match the attributes in the N.

For complex matching, in order to simplify the matching process, First M: N matching convert to M: 1 match and 1: N matching, and then get M: N matching. For example: local query interface with attributes "publication", which contains two condition attributes "date from" and "date to" property, when search ontology conceptual model,these has two classes "from" and "to" in the ontology, and the main class of class attribute "from" and "to" in the ontology concepts is the "publication date", then we can conclude that the concept is similar by comparing the "publication" and "publication date",thus can deduce "publication" and "from", "to" as the 1: M matching, "date from" and "publication date" as the M: 1 match, "from" , "to" and "date from", "date to" as the M: N matching.

## 3.3. Integrated Query Interface

Deep Web Integration Query interface provides a method to access the web database which belongs to a particular area,it can be seen as many local query interfaces with the similar field over the global view.According to observations,with the same field and the gradual increase in available resources, the size of query interface model to stabilized[5], and through the research found the user query during the course of the average number of parameters to fill 3-5.Accordingly this paper, an integrated query interface constructed based on similar domain to pick out the deep web entry form n+ 2 of the most important query parameter is set to Integrated Query Interface parameters (n is the different areas of the average number of parameters from the user used),and then filling the attributs using the domain

ontology,and improve property values of integrated query interface. The integration of thought are as follows:

### 3.3.1. determine the properties of labels and types of integrated query interface.

Because the different deep web interface each are not identical, and the query attribute name is not the same, so this paper the attributes of integration query interface are set to the highest class of the concept which are matched the domain ontology,while retaining an matching relation between integrated query interface and the local query interface,and put it into attribute matching table. The local query interfaces based on the observed significance of the property with the same type are not identical,then recorded the number of attributes which synonymous with each local query interfaces,and put the most times attribute types as querying type of integration query interface.

### 3.3.2. determine the value of an integrated query interface attributes.

In the property values of integrated query interface selection process, if the local query interface attribute values matching a conceptual model of ontology,then we can regard as the attributes of the local query interface and the concept of ontology matching. If the property value is not in the concept of ontology-owned property values, and with the value of all instances of the concept does not match,then the properties of the local property values as a new instance of the concept to the ontology [6].

In this paper, the process of establishing global query interface, the use of ontology technology to determine the interface properties and property tags, and in order to ensure important attributes in front of the position of the global query interface,calculating the frequency to determine the attribute order. The matching number of interface attributes based on the counter in the ontology,if the higher the value,the more important concept it is. Through the records the degree of important querying interface properties,will descending sort order by the matching number in the ontology,thus ensuring the attributes in front of the local query interface location which in global query interface is also the front.

## 3.4. Attribute Extractor

Deep web query interface is a HTML page which begin <form> to </ form> and it can receipt user enters to retrieve local databases. The attributes of those form was constructed by query elements and form tags,the query elements was used to be receipt user query conditions,those generally include input text box and drop-down menu such as textbox,checkbox,radio button and selection etc. Form label is referred to explain the query input such as Html tag label andsome text information around the input which also to explain the input. In order to correctly understand the information of the query interface must be accurately extract the every query elements.Therefore,the pattern of the query interface can be expressed as F=(A,S),where A is the Action in the html form, S is the information in the form, and it can expressed as S={S1,S2……Sn-1,Sn-2}.each data Si can be expressed as Si=(P,T),In the equation, P represent form query element,and T is the tag.As shown in Fig2,the query interface is coded by html format,and the tags are appeared in pairs.so the think of the attribute extraction as follow[7]:

1)first scan the HTML code,found the form tag,extracted the form method,action,and put those information deposit to A binary group F, recorded as A=(method=post,action=subSearch.action）.

2)Then check the information in the form tag,if the tag is <input>,check the value of the type inner the input tag,and if the the value is not the submit or button,extract the name and value,and deposit into the S of the binary group F,recorded as S=(input,(name=title,value=null)) ， if the value of the type is submit or button,thenS=((input,(value=search)).

3)If the label is <lable>,then standardize the attribute information inner the label tag,such meaningless content as remove ":","()" and so on,and recorded as S=(lable,the standardized content ).

4)If the label is <button>, then S = (button,value = search).

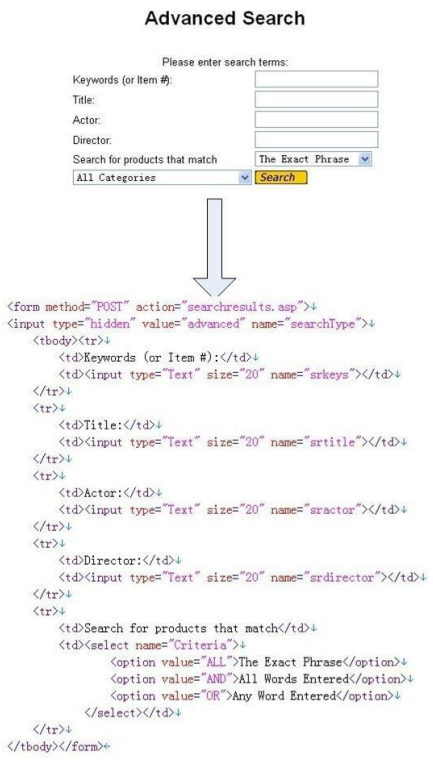


Fig. 2.System framework

Through the above steps,the query interface elements in the html form inductive analysis,parsing each logical attribute name, type, labels, and range to construct the overall model of the query interface,and to achieve deep web query interface mode automatically extract.

## 3.5. Query Translation

Query translation is essentially a process of the query from integration interface to local interface.During the query translation,in this paper,using domain ontology,attribute matcher and attribute matching table allows users to query every deep web databases[8].

The query translation machine inclueded attribute matcher and attribute matching table,attribute matcher is responsible for integrationed query interface to local interface the conversion,and the mapping relationship of the properties of the query interface was stored in attribute matching table.In this paper,the process of establishing an integrated query interface at the same time to complete attribute matching table initialization. When the user queries in the integrated query interface, the first match in the attribute matching table to find whether there is a match, in case the attribute is matching then convert it to the local interface directly,if cannot match is using attribute matcher and domain ontology to attribute transformation and conversion to local interface.

As a result of the heterogeneity of the local query interfaces, the model of query interface must be inconsistency,which makes the query condition from integration query interface to different local interface must be rewrite.For example, in different local query interfaces on the "price" of the query, some are text tag, and some are select tag.This makes the process of converting the query to find the closest conditions,that is,finding the minimum coverage of initial query.However different local query interface has different query attribute,so the solution to different types of attribute as followed:

text box type,if the target query interface has a text box type, which can directly put the user query conditions to local query interface,without the need for special Treatment.

Text list box type, if the target query interface is text list box type,such as select tag,which need to compare user entered query condition with the value of the list box type,then select the highest value as a local interface query condition.

# 4. Experimental results

in this paper,the method mainly based on recall and precision of the two indicators to be assessed.R represent the real exist result in data set,N represent the results obtained on the experiment, then $R \bigcap N$ represent the correct result.so the recall and precision mathematical formula is as follows:

$$\mathbf{Precision}=\frac{|R \bigcap N|}{|N|}\times 100\% \quad (1)$$

$$\mathrm{Re}call = \frac{|R \cap N|}{|R|} \times 100\% \quad (2)$$

According to the formula shows that a high recall rate system, the precision rate cannot meet demand, on the contrary, high precision, the recall rate of less than practical applications. In order to assess the performance of the system,the need to calculate the F-measure,which is recall and precision of the weighted average.A higher F-measure value indicate that recall and precision have a higher value.

$$F_{-measure} = \frac{2 \times \mathrm{Precision} \times \mathrm{Re}call}{\mathrm{Precision} + \mathrm{Re}call} \times 100\% \quad (3)$$

In this paper,using the Book field and movie field in TEL-8 dataset to test the method,the experimental results shown as table1 and table2:

Table 1.query translation result in BOOK field

| Domain | Interfaces used | Interface Labels | Extracted label | Right labels | Recall | Precision | F-measure |
|---|---|---|---|---|---|---|---|
| Books | 20 | 379 | 346 | 314 | 82.85% | 90.75% | 86.62% |
| Books | 40 | 694 | 627 | 583 | 84.01% | 92.98% | 88.27% |

Table 2. query translation result in Movies field

| Domain | Interfaces used | Interface Labels | Extracted label | Right labels | Recall | Precision | F-measure |
|---|---|---|---|---|---|---|---|
| Movies | 15 | 103 | 91 | 87 | 84.47% | 95.60% | 89.69% |
| Movies | 35 | 273 | 251 | 239 | 86.08% | 95.21% | 90.41% |

# 5. conclusion

In the deep web data integration process, the query conversion technology has been a research focus, the query transformation in the course of the study the use of domain ontology and schema match techniques to improve conversion accuracy, and adding an matching map to speed up the matching speed in the query translation module. The experiment results show that the recall rate, precision and F-measure remained at a stable range, the proposed method in this paper is feasible to query translation.

# 6. References


[1] He Bin, Kevin Chen chuan Chang. Statistical schema matching across web query interfaces [C]. Proceedings of the 2003 ACM SIGMOD International Conference on M anagement of Data, SanDiego, California, USA, 2003, 2172 228.

[2] Wu Wen sheng, Clement T Yu, AnHai Doan,Meng Wei yi . Anin-teractive clustering-based approach to integrating source query interfaces on the deep web [C]. Proceedings of ACM SIGMOD International Conference on Management of Data, France, 2004, 952 106.

[3] Neches R ,Fikes R E ,Gruber T R et al. Enabling Technology for Knowledge Sharing.AI Magazine ,1991 ,12 (3) :36～56

[4] Li Xu,David W.Embley:A composite approach to automating direct and indirect schema mappings[J].Inf.Syst.2006,31(8):697-732.

[5] He Bin, Kevin Chen Chuan Chang . Making holistic schema matching robust : an ensemble approach [C]. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 2005, 4292 438.

[6] Wang Ji ying, Wen Ji rong, Fred Lochovsky et al.Instance- based schema matching for web databases by domain-specific query probing[C]. Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, 2004, 4082 419

[7] Jiying Wang,Fred H.Lochovsky.Data Extraction and Label Assignment for Web Databases[C].In Proceedings of International World Wide Web Conference,2003,pp187-196.

[8] Liang Hao,Wanli Zuo,Fei Ren,Junhua Wang.Translating Query for Deep Web Using Ontology[C].In Proceeding of International Conference on Computer Science and Software Engineering.2008,pp427-430.