Describing the Semantic Relation of the Deep Web Query Interfaces Using Ontology Extended LAV

LIANG Hao
Department of Computer Science and Technology, Jilin University, Jilin Changchun, China; Department of Information, Changchun Taxation College, Jilin Changchun, China
Email: Liangh434@163.com

ZUO Wan-Li
Department of Computer Science and Technology, Jilin University, Jilin Changchun, China; Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Changchun, China
Email: Wanli@jlu.edu.cn

REN Fei
Department of Computer Science and Technology, Jilin University, Jilin Changchun, China; Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Changchun, China
Email: Renf854@163.com

***Abstract*—The key element in a Deep Web information fusion system is the data source modeling problem, which is the determinant technical factor of the whole system. The query interfaces provided by the Deep Web are the clues to disclose the hidden schemas. But the complicated semantic relationships in the query interfaces lead to the lower generality and ability of local as view (LAV) method in the traditional information fusion system. An approach of extracting attributes and semantic relationships from the interfaces utilizing Ontology is present in this paper, and WordNet is introduced as an Ontology instrument. The semantic relationships between semantic related attributes are evaluated by the WordNet. The meaningless attributes are instantiated by instance information embedded in the interfaces. A semantic matrix is generated and used to evaluate the semantic related groups in the specific domains. The expression ability of LAV is extended by the mapping and matching mechanism based on the semantic related groups. The experiment is carried out on the famous dataset, and the results show the efficiency of Ontology extended LAV of building mappings between local schemas and mediator schema.**

***Index Terms*—Deep Web, information fusion, query interface, local as view, Ontology, WordNet**

I. INTRODUCTION

After Tim Berners-Lee created the first Web site in August 1991, we have seen the tremendous growth of the Web. The Web influences almost every aspect of people's daily lives by providing information sharing platform. At the same time, an increasing number of databases are becoming accessible through search interfaces, and many of these sources are E-commerce sites supported by databases. These databases are called Deep Web, which can not be crawled by the search engines. The Web has been rapidly "deepened" by massive hidden databases. While the Surface Web has linked billions of static HTML pages, a far more significant amount of information is believed to be "hidden" in the Deep Web, behind the query forms of searchable databases. A survey in April 2004 estimated there were more than 450,000 online databases [1]. Myriad information may not be accessed through static URLs because they are presented as result after users submitted the query. The Deep Web databases require manual query interfaces and dynamic programs to access their contents, thus preventing Web crawlers from automatically extracting their contents and indexing them, and therefore not being included in search engine results [2]. To extend the human physical limitation of information processing in the information age, the information fusion techniques is chosen as the main method.

Information fusion is an interdisciplinary field which research combining and merging the information or data from different information sources [3, 4, 5]. The output of the information fusion system is a unified view which is built on the different heterogeneous data sources. The differences between sources are conceptual, contextual and typographical representations [6, 7].The traditional researchers of information fusion field concentrated their focus on the database community. With the information explosively spreading on the Web, they try to integrate the information on the Web. However, the nature of Web

doi:10.4304/jsw.5.1.89-98

is different from those of traditional databases and multi-databases, as information on the Web are relatively more dynamic, less structural, larger in scale, more open, less controllable, more heterogeneous, more distributed, and more hyper-linked [8].

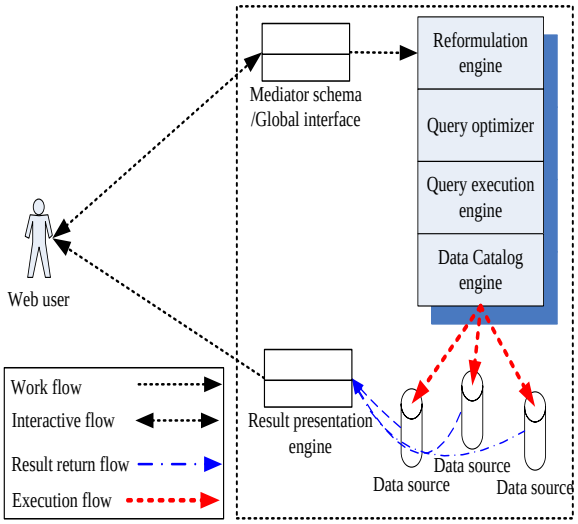

Figure1. Mediator model of information fusion

We always consider Amy's case as the goal of information fusion on the Web, which is about a girl who is moving to new city and she needs something. She has to find a new apartment, a new job and a car. After sources hunting on the net, she must learn details of querying each source, that maybe real-tor.com, cars.com and monster.com. To solve Amy's problem, we need help her find online databases she needs firstly, merge the sources in the same domains, describe the semantic mappings between the schemas of the difference sources and provide a uniform query interface to a multitude of data sources. The information fusion on the web is facing many challenges. The first one is the source modeling description and there are two most major methods to solve the problem: LAV (local as view) and GAV (global as view). In this paper, we present an approach of extending the generality and expression ability of LAV utilizing Ontology technique.

In the context of the Web, we do not differentiate between the terms Web information fusion, Web information integration and Web data integration, and use them interchangeably to refer to the task of combining information on the Web. The query interface and query form are the same meaning too.

II. The Mediator of Web Information Fusion

Kambhampati et al. presented us with three different views of information fusion by including previous studies about the information integration [9]. The first one is a database view, which is mainly about the integration of autonomous structured source data. The second one is a Web service view, which is mainly about combining and composing information provided by multiple web sources. The last one is an information retrieval and natural language processing view, which is about computing textual contents from disparate web text sources. Three research models of information fusion on the Web were presented: the data warehouse model, mediator model and the search model which are corresponding to the three views.

In Fig 1, we present the mediator model of information fusion framework. It is composed of six major functional components. To realize the six components, we are facing some technique challenges such as generating a mediator schema, reformulating the queries, optimizing queries, executing queries, data source indexing, information extracting and result presenting. To generate a mediator schema/global interface is the first step should be realized, during this procedure we need analyze the data sources and introduce a source modeling description mechanism to depicting schemas of local sources and the semantic mappings between the schemas of local sources and mediator schema. Some source information is very necessary [9], which contain following details:

1. Logical source contents, the kinds of objects which are stored in the source such as books, new cars.
2. Source capabilities, if there is a structured query handling mechanism in the source.
3. Source completeness and reliability.
4. Statistics information of the data .
5. Physical properties of the source and network.

To answer queries using the information sources the system needs mappings that describe the semantic relationships between the mediated schema and the schemas of the sources. These mappings and source information are the main component of source descriptions.

Finishing source description between data sources, the next step is how to effectively reformulate a user query into a series of queries which can be executed on each dispersed and heterogeneous data sources. Effective reformulation mechanism must be sensitive to the constraints of data sources and make sure to access the smallest number of most relevant sources when answering the query.

After a posed query has been reformulated, it needs to be executed efficiently. While many techniques of distributed data management are applicable in this case, several new challenges arise, all because of the dynamic and heterogeneous nature of sources [7]. Unlike traditional database community, the execution of data integration system can not be divided into query optimization step and followed by query execution step. There is less information than the traditional database community for the query optimizer. As a result, the optimizer may not have enough information to decide on a good query execution plan, and the result of the query execution plan may be very poor if the sources do not respond exactly as expected. To handle this technique challenge, the efficiency of source descriptions is the basic of the query optimizing.

The schemas of the sources are semantic diversity from each other, for example, the different attributes in different schema may mean the same thing. After user poses a query to the mediator schema, the query is translated into a series of different query segments according to the source descriptions between the mediator schema and sources schemas. The query segments will be sent to different sources which indexed by data catalog engine. At this step of information fusion system, there will be a data catalog engine by adding caching information. The data catalog engine is used to generate a virtual data warehouse, based on which the system to locate the different sources being integrated.

The Web pages are the returned results of Web sources, the structure data representing the schema of the data sources is scattered in the semi-structure Web pages. Much of the information content of the Web is presented in natural language and is organized in a form that is most suited for human intuition. Information extraction (IE) is a type of information retrieval the aim of which is to automatically extract structured information, i.e. categorized and contextually and semantically well-defined data from a certain domain, from unstructured machine-readable documents. A broad goal of IE is to allow computation to be done on the previously unstructured data. A more specific goal is to allow logical reasoning to draw inferences based on the logical content of the input data. It is necessary to use information extraction to extract the scattered clues of the hidden schema.

During all the procedures, the source modeling/description is the most important technical detail. It determines the carrying out of the following steps of the information fusion system. There are two exiting approaches to address the source modeling/description, LAV and GAV.

III. THE LAV AND GAV

A traditional data integration application is started with a set of exiting and heterogeneous data sources. Hence, the first step of the application is to design a mediated schema that describes the logical and physical information of sources, and expose the aspects that may be of interest to users. In principle, one could use arbitrary formulas in first-order logic to describe the data sources. But in such a case, sound and complete reformulation would be practically impossible. Hence, several approaches have been explored in which restricted forms of first-order formulas have been used in source descriptions, and effective accompanying reformulation algorithms have been presented. In this paper, we describe two of such approaches: the GAV [10, 11, 15, 16] and LAV [12, 13, 14, 15, 16]. We would like introduce the GAV and LAV as the view of Maurizio [17].

We can formalize a data integration system I in terms of a triple<G, S, M>, where G is the global schema, expressed in a language L_G over an alphabet set which comprises all symbol for each element of G. S is the source schema, expressed in a language L_S over an alphabet set which includes all symbol for each element of the sources. M is the mapping between G and S, constituted by a set of assertions of the forms $q_S \cong q_G$ and $q_G \cong q_S$ where q_S and q_G are two queries respectively over the source schema S and global schema G. Queries qS are expressed in a query language L_M, S, and queries qg are expressed in a query language L_G, M. In intuitively view, an assertion $q_S \cong q_G$ specifies that the concept represented by the query q_S over the sources corresponds to the concept in the global schema represented by the query q_G.

A. GAV

In the GAV, the mapping M associates to each element g in G and a query q_S over S. A GAV mapping is a set of assertions in the form of $g \cong q_S$. From the modeling point of view, the GAV approach is based on the idea that the content of each element g of the global schema should be characterized in terms of view q_S over the sources. In some sense, the mapping explicitly tells the system how to retrieve the data when user wants to evaluate the various elements of the global schema. This idea is effective whenever the data integration system is based on a set of sources that is stable.

B. LAV

In the LAV, the mapping M associates to each element s of the source schema S and a query q_G over G. A LAV mapping is a set of assertions in the form of $s= q_G$. From the modeling point of view, the LAV approach is based on the idea that the content of each source s should be characterized in terms of a view q_G over the global schema. This idea is effective whenever the data integration system is based on a global schema that is stable and well-established in the organization. A notable case of this type is when the data integration system is based on an enterprise model, or Ontology [18].

C. Benefits of LAV

The approach of LAV is presented after GAV, the modeling point of view of LAV provide more flexibility than GAV. When describing information sources, it becomes easier because we do not need to involve knowing about other information sources and all the semantic relationships between each of sources. As a result, when designing a data integration system, it is easy to accommodate new sources, which is particularly important in the applications of involving hundreds or thousands of sources. The descriptions of the information sources could be more precise than GAV. The source description is greatly depending on the expressive ability of the view definition language, but it is easier to describe precise constraints on the contents of the sources and sources in relational structures. The ability of describing the constraints or relationships between the sources is very important, because the source description provide the mechanism for the follow up works. It is necessary to use right modeling view of source description to provide on the fly query optimizing method and make sure to select a minimal number of sources relevant to carry out the query.

Ⅳ. AUTOMATICALLY ATTRIBUTE EXTRACTION FROM DEEP WEB QUERY INTERFACES USING WORDNET

While the Surface Web has linked billions of static HTML pages, a far more significant amount of information is believed to be "hidden" in the Web databases, which is named Deep Web [1, 2, 19, 20, 21]. The search engines or Web crawlers cannot access Deep Web directly. Most of the Deep Web can only be accessed through dynamic query interfaces which contain HTML form elements. In actual application, it is even more concerned about the contents of Deep Web. The reason is not hard to understand, this part of the structured data is more meaningful to be integrated and more technology can be used [7]. A large number of E-commerce Web sites provide small amounts of Deep Web information in answer to user queries. Finding the relevant E-commerce sites and accessing, retrieving and indexing the huge amount of Deep Web information raise challenging research issues. The query interfaces is the "shadow" of the schemas hidden behind the Surface Web, the results returned from the sources are the clues of the sources schemas. As we all know, the results are a kind of dynamic responses to the query instances that the users posed against the query interfaces. Instead of carrying out research on the templates of the results, we would like to parse the query interfaces to disclose the hidden schemas. To generate a more general view for the information fusion system of Deep Web, we are facing some technical challenges. They are described as following:

S1: The same word in different schemas of query interfaces often has the same semantics, maybe the formalizations are changed.

S2: The words of two different forms are synonyms. This requires the use of thesaurus or dictionaries. In many cases, some special subjects like biology and philosophy are needed specific domain dictionaries.

S3: There are many hierarchy relations between the words, e.g., A is a hypernym of B if B is a kind of A, on the other side B is hyponym of A. For example, Car is-a Vehicle it means Vehicle is hypernym of Car, Automobile is-a Vehicle too. Thus we can get that Car is equal to Automobile semantically.

In order to understand the interface semantically, we introduce the semantic layer to deal with the Semantic Web problems which defined by Tim Berners-Lee [21]. The primary Ontology instrument introduced in this paper is the WordNet [23, 24] with the aim to get the semantically understanding of the information in the interfaces and mapping attributes. The automatically attributes extraction algorithm is described detailed in Ref [25].

Ⅴ. EXTENDING THE EXPRESSION ABILITY OF LAV BY ONTOLOGY

During the procedure of generating the mediator schema, we can extend the ability of expression of LAV. The query interface is the "shadow" of the hidden schema. The ability to reflect the true situation of the hidden schemas is conditional upon the description clarity of the "shadow". Before using the LAV to describe the semantic relation and mapping between the mediator and each source, the generality of LAV is the big challenge of the information fusion system. We note that the LAV approach favors the extensibility of the system: adding a new source simply means enriching the mapping with a new assertion, with other changes [17].The extensibility is very important when integrating hundreds and thousands of the Deep Web sources. However, the nature of Web is different from those of traditional databases and multi-databases, as information on the Web are relatively more dynamic, less structural, larger in scale, more open, less controllable, more heterogeneous, more distributed, and more hyper-linked. Each query interface is different from each other even in the same domain. So the workable way is to build the semantic mapping relation between the mediator schema and the "shadow" of the Deep Web sources which expressed in the form of query interfaces. The query interface is composed of attributes and the semantic relationship between attributes. Based on the extended attributes extracted from the query interfaces, the ability of expression of LAV can be extended and improved.

We take the Book domain in UIUC dataset [25] as a target of Deep Web information fusion system. The attributes of the query interfaces can be extracted by the algorithm presented above. We build a mediator schema Book (title, author, category, ISBN, publisher, publishdate, binding) based on the typical and well designed query interfaces. There are two views about the local schemas in dataset:

V_1 (bookname, surname, subject, ISSN)

V_2 (title, firstname, lastname, topic, UPC).

V_1 contains all the history books published since 1960 and V_2 contains computer books published after 1970. When integrating the data sources represented by V_1 and V_2, the semantic difference between the attributes of mediator schema and local schemas should be matched first. To handle this challenge, we can use the output of the procedure of attribute extraction, which contains the semantic mapping information between the attributes showed in query interfaces. We can treat the output as an original domain specific Ontology instead, which is used to share the consistent and formal information in a specific domain.

Definition 1: Given two words w_1 and w_2, if w_1 is semantic related to w_2 in domain Ontology O and w_1 is semantic reachable from w_2 by semantic relation r_1, it is represented as $w_1 \overset{r_1}{\Rightarrow} w_2$.

Definition 2: If the semantic relation between concepts in O is symmetrical and $w_1 \overset{r_1}{\Rightarrow} w_2$, there is a semantic relation in O confirms $w_2 \overset{r_2}{\Rightarrow} w_1$, the relation between r_1 and r_2 is symmetrical.

Definition 3: Given two different words w_1 and w_2, if there is a direct same ancestor word w_3 in the WordNet, then w_1 and w_2 are in the same synset.

Take the WordNet as an example, it is easy to find that "name" is the hypernym of "title" and "title" is the hyponym of "name". Here we get $name \overset{hyponym}{\Rightarrow} title$ and $title \overset{hypernym}{\Rightarrow} name$, the relation "hyponym" and "hypernym" is symmetrical to each other. During the extraction of the attributes in the query interfaces, the semantic relationships are treated as important as concepts, because the relationships bridge between the concepts. It is easy to find that $bookname \overset{hyponym}{\Rightarrow} title$ and $title \overset{hypernym}{\Rightarrow} bookname$ in the Book specific domain Ontology.

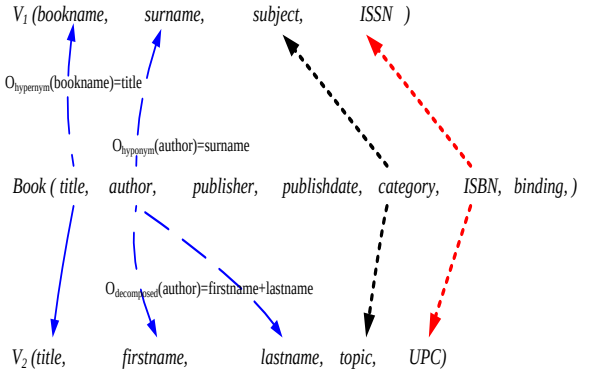

Figure 2.Example of mapping between local views and mediator view

In traditional information integration, the relation between the local views and mediator schema is described directly by LAV, but the query interfaces build a shadow layer between the mediator schema and local views. During mapping between mediator schema and local views, there is a translating procedure. In Figure 2, it shows an example between Book, V_1, and V_2. The semantic relationships between "title" and "bookname", "author" and "surname" ,"firstname", "lastname" are easy confirmed. There are still two additional situations to be dealt with: the first one is the method by which we can calculate the semantic relation between the semantic related attributes "category", "subject" and "topic"; the second one is how to confirm the semantic relationship between some meaningless or semantic unrelated attributes, such as "ISSN", "ISBN", "UPC" in Fig 2.

A. The calculating measure of semantic related attributes

To handle the closely semantic related attributes situation, we can use the WordNet which provide the measure to calculate the semantic distance between any two words. Instead of using frequency of occurrence as an index of familiarity, WordNet uses polysemy. If an index value of 0 is assigned to words that do not appear in the dictionary, and if values of 1 or more are assigned according to the number of senses the word has, then an index value can be made available for every word in every syntactic category.

A simple example of how the familiarity index might be used is shown in Table 1.The super ordinates of bronco are requested, WordNet can respond with the sequence of hypernyms shown in Table 1.The hypernyms of bronco include simply: bronco @-> pony @-> horse @-> animal @-> organism @-> entity. This shortened chain is much closer to what a layman would expect. The index of familiarity should be useful, therefore, when making suggestions for changes in wording. A user can search for a more familiar word by inspecting the polysemy in the WordNet hierarchy. WordNet organizes nouns, verbs, adjectives and adverbs into synonym sets, which are further arranged into a set of lexicographers' source files by syntactic category and other organizational criteria. Some detailed description is in [24].

TABLE I. THE HIERARCHY EXAMPLE OF BRONCO

Word	Polysemy
bronco	1
@-> mustang	1
@-> pony	5
@-> horse	14
@-> equine	0
@->odd-toed ungulate	0
@-> placental mammal	0
@-> mammal	1
@-> vertebrate	1
@-> chordate	1
@-> animal	4
@-> organism	2
@-> entity	3

We present a method to calculate the semantic distance in WordNet. The $S(w_1, w_2)$ presents the semantic distance between two different words w_1, and w_2, the $Len(w_1, w_2)$ is the length of the shortest path between w_1, and w_2 in WordNet, the $turns(w_1, w_2)$ presents the times of turn on the shortest path between w_1, and w_2. The δ and ε are the experimental values, the value domain of δ, ε is [0.3~0.5] and [0.5~0.7]. In this paper, the semantic values are calculated with δ=0.5, ε=0.7. The smaller the $S(w_1, w_2)$ the distance is further between the w_1, and w_2.

We can define a threshold α to evaluate the semantic distance by manual. The value of α is an experimental, it is can be set differently when dealing with different sources. The value of *S (category, subject)* is 0.71, which is the same as the value of *S (category, topic)* calculated in WordNet. However, this situation is not a coincidence, because we can find that "topic" and "subject" are in the same synonym set in WordNet, in another word the value of *S (subject, topic)* is 1.0.

$$S(w_1, w_2) = \begin{cases} \dfrac{1}{\delta * len(w_1, w_2) + \varepsilon * turns(w_1, w_2)}, if \quad w_1 \\ and \quad w_2 \quad are \quad not \quad in \quad the \quad same \quad synset \\ 1, \qquad else \end{cases} \quad (1)$$

B. The calculating measure of meaningless or semantic unrelated attributes

There are some useful words such as "From", "To" which we can not get any description WordNet, but this kind of meaningless words do play a supporting role in the query interfaces of Airfares, .etc. To handle this situation, we can use the results of element text extraction [25]. During the element text extraction, some instance information is extracted from the query interfaces, this kind of instance information can instantiate the related attributes by attaching some semantic description. When instantiating some attributes, to find correct and related instance information is the key step. In the query interface, the layout format includes the information describing the relation between the each labels and elements and free texts and the information can be used as a kind of heuristic information. We consider the texts of instance information in the query form and compute the visual distance with respect to each field of the form. We order the instances into a list and choose the top one to instantiate the related attribute.

The main reason why Deep Web can not be access directly by search engine is that the data in the hidden database is presented as result after users submitted the query. Especially in E-commerce websites, there are detailed data model and data describing the information of diverse commodities. By extracting the data model and data, it is easy to find related instance information of the abbreviations or acronyms attributes. There are three main methods of extraction: manual approach, wrapper induction, automatic extraction. The results of different approaches are different. There is some detailed introduction by Bing Liu [26]. In paper, we use a kind of heuristic automatic extraction method discussed in Ref [25].

The meaningless or semantic unrelated attributes are widely used in the query interfaces of Deep Web, where they are used to represent public and consistent concepts, such as "ISSN", "ISBN" and "UPC" in the Books domain. Take "ISBN" as an example, it is a series of numbers between 0-9, which means "International Standard Book Number" in the Book domain. The frequency of using "ISBN" to search for a book is rarely low. However, the "ISBN" can identify a book uniquely, but it is more difficult to remember a series of numbers than some words in the human sense. So we focus on the general attributes represented by words and do no further research on the meaningless or semantic unrelated attributes except to instantiate them utilizing some instance information extracted form the query interface.

C. Generating the semantic matrix and candidate semantic groups

The *SM* is short for semantic matrix. The semantic distance is calculated by method discussed above. The matrix is an upper triangular matrix, each cell of which is the semantic distance *S (word1, word2)*. The algorithm of generating the semantic matrix is shown in Fig 3. The block upper triangular matrix in the *SM* can be used to evaluate the semantic groups. There are two parameters in the *GSM* algorithm, the α is used to filter the matrix and β is the number of the cell which semantic distance is 0 after filtered by α in the partial upper triangular matrix. The value of β is a kind of inching switch value in aim to deal with the zero semantic distance during the *GSM*. However, in the all the domains of Deep Web, the number of attributes in the semantic matrix is very limited, so the value domain of β is [1, 2] experimentally.

In Fig 4a, it is the original semantic matrix and the semantic matrix with α=0.5 is in Fig 4b. It is easy to find that the whole domain is only one semantic group when α=0. By adjusting the value of α and β, different clustered semantic groups can be generated from the semantic matrix. The example on the semantic matrix with threshold α=0.5 and β=1 in Books domain is shown in the Fig5.

```
Algorithm: Generate Semantic Matrix (α, β)
1:   MaxS=0;
2:   for (i=0; i<n; i++)
3:     for (j=i; j<n; j++)
4:       {RS=GetS (ai, aj);   //GetS (ai, aj) is a method of
       evaluating the semantic distance in the WordNet, ai is a
       word.
5:        if (MaxS<RS )
6:          { MaxS=RS; S= ai; ai=aj; aj= S;}} // reorder the
       word list by semantic distance.
7:   To generate the Semantic Matrix use the GetS values. The
     SM is a block upper triangular matrix.
8:   Set a value for threshold α.
9:   for(i=0; i<n; i++)
1     for(j=0; j<n; j++) {
^1      if (SM[i, j] < α)
.1        SM[i, j]=0; } // filter the SM[i, j] by threshold α.
^1   Generate candidate semantic groups by a reverse upper
3:   triangular matrix processing procedure.(RUTMP)

Algorithm: RUTMP (SM) // SM:Semantic matrix
1    i= column number;
2    if (i==0)
3      generate final semantic group
4    else if (Reverse (SM[i]) and Zero(SM[i])< β)
       // the function of Zero is to calculate the number of zero
     cells, if the number of zero cells in the partial upper
     triangular matrix is more than β, the algorithm begins a
     backtracking procedure.
5      RUTMP (SM[i-1])
6        else generate temp semantic group and cut the
     semantic matrix.
```

Figure3. Algorithm of generate semantic matrix and groups

After extending the expression ability of LAV, the local views can be described as Fig 6 by the assertion generated by the Ontology extended procedure based on

the semantic groups in the query interfaces. The semantic relationships we concerned are 'hypernym', 'hyponym', 'instance of' and some description based on domain knowledge such as the author is equal to firstname and lastname.

S1: *V_1 (bookname, surname, subject, ISSN) ⇒ Book (title, author, category, ISBN, publisher, publishdate, binding)* ∧(publishdate >1960) ∧(category= "History") ∃ ($O_{hypernym}$(bookname)=title, $O_{hyponym}$(author)=surname, $O_{instanceof}$(category)=subject)

S2: *V_2 (title, firstname, lastname, topic, UPC) ⇒ Book (title, author, category, ISBN, publisher, publishdate, binding)* ∧(publishdate >1970) ∧(category= "Computer") ∃ ($O_{decomposed}$(author)=firstname+lastname, $O_{instanceof}$(category)=topic)

VI. Experiments and Discussion

The data set was downloaded from the UIUC web integration repository [27]. This dataset contains the original query interfaces and their manually extracted query capabilities of 447 deep Web sources from 8 representative domains. It contains airfares (49), automobiles (97), books (67), car rentals (25), hotels (39), jobs (52), movies (78), and music records (70).

We selected Books (40), Movies (25), Automobiles (25), Hotels (30) and Airfares (35) domains in the dataset. The interfaces of Movies are the simplest interfaces but with representative meaning, there are less attributes and semantic restriction in the query interfaces and the Music Records is in the same situation. The query interfaces of Airfares and Automobiles domains are the most complicated in the whole dataset, because there are so many attributes with kinds of semantic restrictions and more form controls. So we decided to carry out experiments in the most complicated query interfaces to test our efficiency of the algorithm.

	title	surname	firstname	lastname	author	writer	category	subject	topic
title	1	0.86	0.33	0.25	0.17	0.33	0.78	0.55	0.55
surname		1	0.56	0.15	0.15	0.15	0.33	0.33	0.33
firstname			1	0.67	0.11	0.22	0	0	0.11
lastname				1	0.25	0.12	0.12	0	0.12
author					1	1	0.38	0.73	0.18
writer						1	0.25	0.73	0.18
category							1	0.71	0.71
subject								1	1
topic									1

(a) The original semantic matrix

	title	surname	firstname	lastname	author	writer	category	subject	topic
title	1	0.86	0	0	0	0	0.78	0.55	0.55
surname		1	0.56	0	0	0	0	0	0
firstname			1	0.67	0	0	0	0	0
lastname				1	0	0	0	0	0
author					1	1	0	0.73	0
writer						1	0	0.73	0
category							1	0.71	0.71
subject								1	1
topic									1

(b) The semantic matrix with threshold α=0.5

Figure 4. An example of semantic matrix in Books domain

	title	surname	firstname	lastname	author	writer
title	1	0.86	0	0	0	0
surname		1	0.56	0	0	0
firstname			1	0.67	0	0
lastname				1	0	0
author					1	1
writer						1

(a) The semantic matrix after first step of RUTMP

	title	surname	firstname	lastname
title	1	0.86	0	0
surname		1	0.56	0
firstname			1	0.67
lastname				1

(b) After second step of RUTMP

	title	surname
title	1	0.86
surname		1

(c) The semantic matrix after third step of RUTMP

Candidate Semantic groups:

SG4= {title, surname}
SG3= {firstname, lastname}
SG2= {author, writer}
SG1= {category, subject, topic}

(d) Candidate semantic groups

Figure5. An example of RUTMP on the semantic matrix with threshold α=0.5 and β=1 in Books domain

We have implemented our algorithms with Eclipse 6.0 and used the Jwnl 1.31 to access the WordNet 2.02. There are the results in Table 2 and some experimental indicators are introduced as follow:

M: The number of the whole domain specific query interfaces in the experiments.

N: The number of the parsed-able domain specific query interfaces by our algorithm.

P: The number of the correct parsed domain specific query interfaces by our algorithm.

Recall: the recall of the algorithm, *N/M*100%.*

Precise: the precise of the algorithm, *P/M*100%.*

O: The average number of attributes in original query interfaces.

E: The average number of extracted attributes by the algorithm.

TABLE Ⅱ .THE RESULTS OF EXPERIMENTS IN FIVE DOMAINS

Domain	*M*	*Recall*	*Precise*	*O*	*E*
Books	40	93%	85%	11	21
Movies	25	95%	87%	9	18
Automobiles	25	88%	87.3%	16	24
Hotels	30	87%	83%	16	32
Airfares	35	83%	85.3%	18	43
Average:	31	89%	86%	14	27.6

Our method can parse the query interfaces effectively, but there are still some complicated query interfaces, which are composed by irregular attributes named mechanism and graphic form controllers, can not be parsed.

The aim of extending the expressive ability of LAV is to find the semantic groups in the specific domains firstly, because the semantic groups are the related the attributes which are depicting the specific object markedly. Take the Books domain as an example, semantic groups of it are "author", "title", "category", "ISBN", "publish". "ISBN" is a high-frequency abbreviation shown in all the query interfaces of the domain, so it is statistics result for that "ISBN" is in the semantic groups. The other four are the extracted results by the method. The "category" represents the "subject" and the "topic", because "category" is in the higher level than the other two in the WordNet; the "publish" represents the "publishdate" and "publisher", both of which are derived from the "publish". The "author" is composed of "firstname" and "lastname" and the "title" is the most generic word which depicts the subject information of books.

In the Fig 5, the SG4= {title, surname} is not a steady semantic group in the example, because the "surname" is used to depict the author's information in the query interfaces, but the semantic relation between "title" and "surname" is very close in WordNet when they are in the meaning of "the name of a work of art or literary composition". From the example we can see that is better not to use "surname" in the query interface accompanied by "title", "firstname" and "lastname", because "surname" establishes a kind of close related semantic relation between the SG3 and SG4. The mapping and matching between the mediator schema and local views of the Deep Web in LAV is by the assertion generated on those semantic relationships of semantic groups in the Ontology.

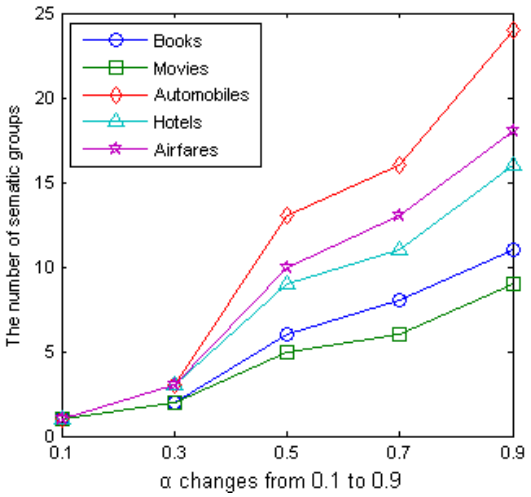

Figure 7.The semantic groups generated by different value of α

We can find that the number of semantic groups is related to the number of attributes used to describe the special object. The more clear the meaning of the words the less different words are used to depict the attributes. It is related to the semantic meaning of the attributes, take the "author" as an example, if everyone knows it means the one who has written the book, so the "writer" will not be showed in the query interfaces of Books domain. *SM* is an effective method to evaluate the semantic expression in the query interface. With the value regulating of α and β, the number of semantic groups generated by the upper triangular matrix algorithm is changing. So the values of α and β are experimental values according to the situation of specific

1 Available at http://sourceforge.net/projects/jwordnet/
2 Available at http://wordnet.princeton.edu/

domain and dealing standard. In the Fig 7, it shows the numbers of semantic groups with the regulating α and β=1. During the experiments, we find the semantic groups generated by the algorithm are effective for describing the semantic relationships when α is 0.5. When the value of α reaches 0.7, the semantic groups are too dispersed to depict the semantic relations between attributes in the query interfaces.

Ⅶ. CONCLUSION AND THE FUTURE WORK

In this paper, we present a new approach of data source modeling description in the Deep Web information fusion system by extending the expression ability of LAV utilizing Ontology. With the aim to get the semantically understanding of the information in the interfaces and mapping attributes, we utilize WordNet to extend the candidate attributes extracted from the query interfaces. The calculating measure of semantic related attributes presented in this paper is based on the structure of words in the WordNet. We also use the instance information in the query interfaces to instantiate the related attributes in order to make some meaningless words or abbreviations meaningful. We focus on the general attributes represented by words, so we do no further research on the meaningless or semantic unrelated attributes. The semantic groups are generated by the upper triangular matrix processing algorithm in the semantic matrix composed of semantic distances of each pair of words in the specific domains. By adjusting the threshold values of α and β, different clustered semantic groups can be generated from the semantic matrix. The generality and expression ability of LAV is extended by building assertion of mappings between local schemas and mediator schema based on the generated semantic groups. In the near further, we will concentrate on reformulating the queries which is posed onto mediator schema by users.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No.60373099, No. 60873235; the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 200801830021; Technology Development Program of Jilin Province of China under Grant No.20070533, No.20080318.

REFERENCES

[1] K. C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang, "Structured databases on the web: Observations and Implications", *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, Paris, France, vol. 33, pp.61–70, September,2004

[2] UC Berkeley, "Invisible or Deep Web: What it is, Why it exists, How to find it, and its inherent ambiguity", Available at www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html, July 2006.

[3] C. Batini, M. Lenzerini, S.B. Navathe, "A comparative analysis of methodologies for database schema integration", *ACM Computing Surveys*, vol. 18, pp. 323–364, 1986.

[4] B.V. Dasarathy, "Decision fusion strategies in multisensor environments", *IEEE Transactions on Systems Man and Cybernetics*, 1991, 21 (5): 1140–1154.

[5] A.P. Sheth, J.A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases", *ACM Computing Surveys*, vol 22 (3), pp. 183–236, 1990.

[6] A.L. Buczak, R.E. Uhrig, "Hybrid fuzzy-genetic algorithm technique for multisensory fusion", in *Information Sciences*, 1996, 93 (3–4):265–281.

[7] A.Y. Halevy, A. Rajaraman, J.J. Ordille, "Data integration: the teenage years", *In Proceedings of the 32nd International Conference on Very Large Data Bases*, Seoul, Korea, ,pp. 9–16, September, 2006

[8] K. Chen, "Large-scale deep web integration: exploring and querying structured data on the deep Web", *Tutorial at In Proceedings of Proceedings of the ACM SIGMOD International Conference on Management of Data 2006*. Chicago, Illinois, USA, June 27-29, 2006, Available: <http://www-sal.cs.uiuc.edu/kcchang/talks/webitutorial-sigmod06-kcchang-jun06.ppt>.

[9] S. Kambhampati, C.A. Knoblock, "Information integration on the Web", *Tutorial at AAAI 07*. Available at http://rakaposhi.eas.asu.edu/aaai2007-i3-tut

[10] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, .et al, "The TSIMMIS Project: Integration of Heterogeneous Information Sources", I*n Proceedings of the 10th Meeting of the Information Processing Society of Japan*, Tokyo, Japan ,pp. 7-18, October 1994.

[11] S. Adali, K. S. Candan, Y. Papakonstantinou, V. S. Subrahmanian, "Query Caching and Optimization in Distributed Mediator Systems", *In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, Montreal, Quebec, Canada, pp. 137-148, June, 1996.

[12] Chung T. Kwok, Daniel S. Weld, Planning to gather information, *In Proceedings of the AAAI Thirteenth National Conference on Artificial Intelligence*, Portland, Oregon, USA, pp. 32-39, August, 1996.

[13] Alon Y. Levy, Anand Rajaraman, Joann J. Ordille, "Querying Heterogeneous Information Sources Using Source Descriptions", *In Proceedings of 22th International Conference on Very Large Data Bases*, Bombay, India, pp. 251-262, September, 1996.

[14] M.R. Genesereth, A.M. Keller, and O.M. Duschka. Infomaster, "An information integration system". *In Proceedings of 1997 ACM SIGMOD International Conference on Management of Data*, Tucson, Arizona, pp. 539–542, May, 1997.

[15] Marc Friedman, Alon Levy, Todd Millstein, "Navigational plans for data integration", *In Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Orlando, Florida, pp.67-73, USA, 1999.

[16] Li Xu, David W. Embley, "Combining the Best of Global-as-View and Local-as-View for Data Integration". I*n Proceedings of 3rd Information Systems Technology and its Applications*, Salt Lake City, Utah, USA, pp. 123-136, June, 2004.

[17] Maurizio Lenzerini, Data Integration: "A Theoretical Perspective". *In Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, Madison, Wisconsin, USA, pp. 233-246, June, 2002.

[18] Thomas R. Gruber, "A translation approach to portable Ontology specifications", in *Knowledge Acquisition,* vol. 5, 1993, pp.199-220.
[19] Singh, Munindar P. "Deep Web structure", in *IEEE Internet Computing*, vol. 6,Los Vaqueros, USA, 2002, pp.4-5.
[20] T.M. Ghanem, W.G. Aref. "Databases deepen the Web", in *Computer*, vol 37, Los Alamitos, CA, USA, 2004, pp.116-117.
[21] M. K. Bergman, "The Deep Web: Surfacing Hidden Value". Available at http://www.brightplanet.com/resources/details/deepweb.html, May 2006.
[22] Berners-Lee, T., Hendler, J. and Lassila, O, "The Semantic Web". *Scientific American*, 2001, pp.34-43.
[23] G. A. Miller. "WordNet: A Lexical Database for English", *in Communications of the ACM*, vol 38, New York, NY, USA, November 1995, pp.39-41.
[24] Christiane Fellbaum, "WordNet: An electronic lexical database", Cambridge, MA: MIT Press, 1998.
[25] Hao Liang, Fei Ren, Wanli Zuo, et al, "Extracting Attributes from Deep Web Interface Using Instances", *In proceedings of 2009 World Congress on Computer Science and Information Engineering*, March-April, Los Angeles, Anaheim, USA, 2009. In press.
[26] Bing Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", *Springer*, 2006.
[27] K. C. Chang, B. He, C. Li and Z. Zhang. "The UIUC web integration repository". Computer Science Department, University of Illinois at Urbana-Champaign, 2003.

LIANG Hao was born in Jilin of China on 16th Jun 1980. He is a Ph.D. candidate at the Department of Computer Science and technology, Jilin University. His current research interests include Web Intelligence, Ontology Engineering and Information Fusion.

ZUO Wan-Li was born in Jilin of China in Dec 1957. He is a professor and doctoral supervisor at Department of Computer Science and technology, Jilin University. Main research area covers Database Theory, Machine Learning, Data Mining and Web Mining, Web Search Engines, Web Intelligence.

He was as a senior visiting scholar, engaged in collaborative research in Louisiana State University (USA) from 1996-1997. He was principle responsible member of 3 national NSFC programs. More than 40 papers of him were published in key Chinese journals or international conferences, 8 of which are cited by SCI/EI. Three books were published by him in Higher Education Press of China and he obtained 3 national and departmental awards.

He is a member of System Software Committee of China's Computer Federation, prominent young and middle-aged specialist of Jilin Province.

Ren Fei was born in Inner Mongolia of China on 14th Aug 1981. She is a Ph.D. candidate at the Department of Computer Science and technology, Jilin University. Her current research interests include Data Mining, Intrusion Detection, Information Security, and Artificial Immune Algorithm.