Integrating Heterogeneous Databases Based on Semantic

Abstract—With the development of information technology, databases, search engines and other information technologies have been widely adopted in the health care system. However, sharing of those information resources is impeded by the heterogeneity and distribution among the independently designed and maintained information resources. Aiming at this problem, we bring forward an Integration of Heterogeneous Databases based on semantic web ontology, which is proposed to draw and represent the common interests of specific groups of people, and hence provide more meaningful and exhaustive query results. As an instance, we developed a unified Child Psychiatry Neuroimaging Portal based on the medical database resources such as the Internet Brain Volume Database (IBVD), Pubmed and so on. A prototype is developed to show that this semantic based integration can work well.

Keywords—semantic, ontology, neuroimaging, integration, heterogeneous databases

I. INTRODUCTION

Along with the constant development of information technology, databases, search engines, and other information technologies have been widely adopted in the health care system. Different hospitals and medical research institutions have accumulated sets of massive medical data, and gradually formed a few authoritative major medical resources search platforms, such as the Internet Brain Segmentation Repository (IBSR), Internet Brain Volume Database (IBVD), famous digital archive of biomedical and life sciences journal literature “PubMed” and more. The main differences between these databases are shown as following.

- Different search conditions: the IBVD require choices in the Drop-down boxes, the IBSR require choose the name of files directly, and Pubmed needs text input.
- Different result formats: the IBVD provides data in tabular format, the IBSR provides images in majority, and PubMed data are mainly in text.
- Different data acquisition methods / technologies: for example, medical images can be captured by X-rays, CT, MRI and other professional methods, and therefore stored in different formats, showing different levels of medical characteristics.
- Different expressions in different levels of medical: different terminologies may indicate the similar or even identical meaning.

II. RELATED WORKS

The above kinds of differences led the sharing of these resources to a real challenge. In fact, in addition to the broad extension of the Internet, few practical solutions can be used to meet the needs of researchers and clinical workers.

Aiming at the above problem, the idea of medical information integration has been mentioned in the past few years. However, neither useful medical integration systems have been realized based on popular authoritative medical databases, nor on advanced image processing and analysis applications. What’s more, as mentioned above, there is a contradiction between the numerous data and traditional URI (Universal Resource Identifier) matching technology. For example, the word ‘body’ in English is ‘kropp’ in Swedish, ‘corps’ in French and ‘telo’ in Czech. Different presentations of the same kind of diagnosis or organ make the information exchange among different countries so difficult that the intelligent diagnosis now only has its influence in a certain area.

To solve the problems mentioned above we set up a system, which based on semantic ontology, of a unified Child Psychiatry Neuroimaging Portal (CPNP) which is supposed to seamlessly retrieve and integrate various kinds of information from available resources of a large number of distributed medical data, for instance, IBVD, IBSR and Pubmed. The semantic web ontology is proposed to draw and represent the common interests of specific groups of people and searching conditions with common attributes, and hence provide more meaningful and exhaustive query results. What’s more, the whole CPNP system exist independently of the kinds of databases like IBVD, which means it need not to know the interface or any other inner information of database under the UI. Without do any change on a certain searching website, the CPNP can provide a common portal in order to make the search easier and more comprehensive.
The structure design of the CPNP system will be introduced in Chapter 1, including the requirements analysis, function modules, and the architectural design. In Chapter 2 we briefly introduce the information integration techniques used in the system, including the idea of applying the semantic web ontology. At last, the system prototyping work that has been done is summarized in Chapter 3.

II. SYSTEM DESIGN AND IMPLEMENTATION

A. The existing Databases

Now, the main medical resources search platforms as following:

IBVD (http://www.cma.mgh.harvard.edu/ibvd/), whose main searching conditions type is the choices in the Drop-down boxes, shows its results, mostly medical examination and diagnosis, in tabular format.

IBSR (http://www.cma.mgh.harvard.edu/ibsr/), whose main searching conditions type is choosing the name of certain files directly, usually result in images and metadata by some compressed files.

Pubmed (http://www.ncbi.nlm.nih.gov/pubmed/), whose main searching conditions type is text, gives back results about diagnosis, treatments and relational papers mainly in text.

 Needless to say, the famous database above each has a large quantity of medical treatments, examination results and diagnosis, which are different from each other. But for the complexity and diversity of the medical data, when user wants to find all the details of a certain treatment diagnosis, he or she has to search in several different databases to collect different kinds of data. It wastes a lot of time and energy to search so many isolated websites.

CPNP is designed to solve this problem. This portal integrated different kinds of databases without changing anything of the databases. With the help of CPNP, user can get different types of diagnosis data only by inputting searching conditions once in the FrontPage. What’s more, with the semantic web ontology, it is easy to draw and represent the common interests of specific groups of people and searching conditions with common attributes, and hence provide more meaningful and exhaustive query results.

B. structure of CPNP

The main structure of CPNP is shown in Fig.1.

①The uniform UI is the browser and website as shown in Fig.2. CPNP FrontPage provides the site introduction, the user manual and the entrance of other modules. Login Module provides the administrators and users login functions. Manager Module has the system administrator operating functions and User Module has the user operating functions. The right subpages show the searching results to users in a required format by tags.

②The Mediator, Data Analysis part in Fig 1, is the most important part, as well as the main work of the author, of CPNP. This part will be introduced in detail in following section.

③The following is the interfaces of databases including IBVD, IBSR and PubMed. In order to improve the portability and decrease the time and energy cost, the whole CPNP system exist independently of the kinds of databases like IBVD. In the bottom of Fig.5 there is a black band between the website and inner databases all searching platforms, which means that the portal need not to know the interface or any other inner information of database under the UI. Without do any change on a certain searching website, the CPNP can provide a common portal in order to make the search easier and more comprehensive.
• Parse key words from a sentence which inputted by the user, and give enough choices to user in order to increase the accuracy of matching.

• Classify the key words by semantic technology: make a word one class including relational words with one or more common attributes.

• Distribute the key words with their classes to the databases by rearranging them in order to fit the input conditions interface of the UI of one platform.

So, we design the structure of Mediator as shown in Fig.3.

D. Semantic based Algorithm for Mediator

As mentioned above, the main goal of Mediator is to deal with the input using semantic technology, and to deliver the classified searching conditions to each database. To better describe the process of the Mediator about how to parse, classify, sort, deliver, summarize and display the data, we define a multidimensional vector as \( S = \{ \text{DB <interfaces_in, interfaces_out, data_type>,Is,Os,Q} \} \). DB dimension inside means the attributes of the platforms in which we search for results such as IBVD and PubMed. The attributes mainly includes the format of input searching conditions, output format and the main data types of its results. Is dimension is the input key elements and Os is the key attributes of one diagnosis. Q is the matching Quality which would be described later.

The input by users is the union of the searching conditions of each database:

\[
\text{Total_Conditions} = \text{DB_A_conditions} + \text{DB_B_conditions} + \text{DB_C_conditions} + \ldots + \text{DB_N_conditions}
\]

(1)

Then the mediator is to complete the matching between the input and the interface of the medical platforms based on the input by user using Description Logic and Semantic Ontology Web Language (OWL). The process of matching mainly takes advantage of the Predicate Logic Inference Machine in order to find the key word in the user input. The key element classes are from the combination of key words and simple string matching.

In order to make the searching results more exhaustive and practical, we define a series of matching level. Each matching level corresponds to a matching quality, which is a real number defined in S, as following:

\[
\text{Matching Level} \in [0,1]
\]

(2)

We can map the distances between the input vector and the vectors in the databases into (0, 1) to represent the Matching Level. The distance between two vectors \( \alpha \) and \( \beta \) can be calculated by the function following function (3), in which \( \alpha \) is the input vector, \( \beta \) is the match vector, \( k \) is a coefficient and \( \eta \) is one of the \( \cup^+ \) union of \( \alpha \):

\[
D = ||\alpha - \beta|| + k||\alpha - \eta||
\]

(3)

The dimensions from A to X is exactly whether the input searching conditions in U0 match the inputs, yet the dimensions from a to x represent whether the conditions in the semantic ontology expanded class U+ match the original input. The coefficient \( k \) is the relational coefficient defined by experiments and practical environments. A more simple way to get \( \alpha - \beta \) is when the ith dimension of \( \alpha \) equal the ith dimension of \( \beta \), then the ith dimension of the error \( \alpha - \beta \) is zero. The relation between U0 and U+ is shown in Fig.5.

The details of the mediator, including the design and implement of each part and the basic algorithm are described in next section.

![Figure 3: The structure of Mediator](image-url)
The pseudo code of match is as following:

```
Get(conditions: a,b,c,d);
//conditions include a,b,c and d.
Match condition A with Database:
{
    if (conditions <= U_element:
        a = U_element_a, b = U_element_b,
        c = U_element_c)
    //Check if match perfectly
    Then Perfect Match, Q = 1; //Q is quality
    Else if (condition <= U+_element
        A,b,c,d ∈ A+_element_contribution)
    // Check if match partly
    Then partly Match,Q = || a - β ||;
    Else then No Match, Q = 0;
}
Sort Results by Q;
Show Results by certain tables.
```

In fact, in medical, especially in the Neuroimaging area, there is one or two elements have a little difference between original conditions and aim conditions would do little influence on the accuracy of matching. It is easy to image that the image and metadata of the brain of a 35-years-old man is almost the same as the brain of a 36-years-old man, for there is little change in the human brain during one’s thirties. As a result, appropriate classify a conditions and make it related to some relational words or congenial elements by simple semantic ontology technology not only can make the results more comprehensive and exhaustive but also can give the user more professional reference to draw a more accurate diagnosis.

After matching, the Mediator shows the results, which is sorted by the Quality, to user by two display choices. The first one is show the results from different databases in different tag pages. The second one is show the results in one page integrated by function (4).

```
Result = R_IBVD ∪ R_IBSR ∪ … ∪ R_Pubmed              (4)
```

III. TESTING RESULTS

A. Testing functions

After getting the input, the mediator parses the key words and classifies them with semantic ontology. This process can help to translate the input to the proper word which can be best distinguished by each database by analysis its interfaces. The total data analysis progress and examples is shown in Fig.5.

In fact, it is recommended to make a feedback during the beginning period to adjust the elements of Semantic Ontology Expand to practice, in order to make the matching result more accurate and comprehensive.
Two test example of the Mediator in Parse Key word and deliver them to different databases is shown in TABLE I below.

<table>
<thead>
<tr>
<th>Query</th>
<th>Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 years old boy with matter in left hemisphere has bipolar disorder</td>
<td>IBVD</td>
</tr>
<tr>
<td>Man in 35 with a disabled leg and his right hemisphere has Affective disorder</td>
<td>Pubmed</td>
</tr>
</tbody>
</table>

Space lacks for a detailed description of it. But from the simple table above we can see that different databases has different input format, and certain conditions, or say key word, are well accepted by different databases such as key word <disabled leg> does not have match diagnosis for IBVD includes mainly data about diagnosis of brain but not body.

B. 3.2 Testing for Semantic based Matching

Compared with the traditional matching ways such like string match, the matching method based on semantic ontology obviously improve the quantity and quality of searching results without any notable influence on the efficiency of search. To the same input, the comparison of search results of traditional matching of search results of traditional matching ways and Semantic based Matching ways is shown in TABLE II.

From TABLE II, we can see that the result searched by traditional method is a subset of that by semantic method. As we all know, the more results doctors get, the more diagnosis and references to use, the more reliable and accurate diagnosis can be inferred from original conditions according to searching results. As a result, the semantic based data analysis can improve both the quantity and quality of computer aided diagnosis, which leads to a remarkable practicability.

C. Time efficiency testing

The process time in TABLE III is the sum of matching time, data transferring time and the time delay under the limit of the bandwidth of local network. From this table we can draw a conclusion that the more time cost by semantic analysis is almost a constant.

IV. Conclusion

In this paper, we set up a system, CPNP, in order to solve the problem of wasting a lot of time and energy to search so many isolated medical platforms and databases. This portal integrated different kinds of databases without changing anything of the platforms. With the help of CPNP, users can get different types of diagnosis data only by inputting searching conditions once in the FrontPage. What’s more, with the semantic web ontology, it is easy to draw and represent the common interests of specific groups of people and searching conditions with common attributes, and hence provide more meaningful and exhaustive query results.

However, as it is an original vision and limited by time and the author’s ability, this CPNP use only a simple level of semantic ontology technology and have a big space to improve in the future. For example, it can be developed by systematic RDF and more sophisticated semantic tools to improve the accuracy of searching. Or we can directly use the interface of database, though it has to do some change on each isolated
databases, but it can increase the search time efficiency obviously.

As a result, the unified Child Psychiatry Neuroimaging Portal (CPNP) based on the medical database resources such as the Internet Brain Volume

Database (IBVD), Internet Brain Segmentation Repository (IBSR) can solve the problem of sharing of the large amount of information resources being impeded by the heterogeneity and distribution among the independently designed and maintained information resources well. We successfully bring forward this semantic based Integration of Heterogeneous Databases the semantic web ontology.

This system successfully draws and represents the common interests of specific groups of people, and hence provides more meaningful and exhaustive query results. This semantic based integration portal is useful not only in information search but also has deeply influence on artificial intelligence.

REFERENCES


