

# Domain Knowledge Extracting in a Chinese Natural Language Interface to Databases: NChiqI\*

Xiaofeng Meng<sup>1,2</sup>, Yong Zhou<sup>1</sup>, and Shan Wang<sup>1</sup>

<sup>1</sup>College of Information, Renmin University of China, Beijing 100872

<sup>2</sup>Institute of Computing Technology, The Chinese Academy of Sciences  
[xfmeng@public.bta.net.cn](mailto:xfmeng@public.bta.net.cn), [swang@mail.ruc.edu.cn](mailto:swang@mail.ruc.edu.cn)

**Abstract.** This paper presents the method of domain knowledge extracting in NChiqI, a Chinese natural language interface to databases. After describing the overall extracting strategy in NChiqI, we mainly discuss the basic semantic information extracting method, called DSE. A semantic conceptual graph is employed to specify two types of modification and three types of verbal relationship among the entities, relationships and attributes. Compared with related works, DSE has more strongly extracting ability.

## 1 Introduction

Natural language interfaces to databases (NLIDBs) is a system that allows the user to access information stored in a database by typing requests expressed in some natural language (e.g. Chinese, English). Since the early 1960s, much of the research on NLIDB has been motivated by its potential use for communicating with DBMS. Now two main issues hinder NLIDBs to gain the rapid and wide commercial acceptance: *portability* and *usability*[1]. These problems are resulted from the system poor ability to cover the domain knowledge in a given application.

Each application domain has different vocabulary and domain knowledge. Such uncertainty of domain-dependent knowledge reduces the usability of NLIDBs in different application domain. Therefore when a system is transported from one domain to another, it must be able to obtain new words and new knowledge. So in order to solve the problem of portability, it must enable the system to extract domain knowledge automatically or semi-automatically at least, in order to build the domain dictionaries for the next step of natural language processing.

Domain knowledge extraction is the process of acquisition of sufficient domain knowledge and reconstruction of domain information especially the semantic and usage of words in order to extend the ability of language processing. Nowadays widely attentions have been paid to studies on knowledge extraction. Researches on computing linguistics proposed some corpus based extracting methods. Technique of domain knowledge extracting from the aspect of NLIDBs is mainly based on database schema. These methods focused on simple extraction of database structure

---

\* This paper is supported by the National Natural Science Foundation of China.

information whereas the potential extraction of semantic knowledge is not fully considered.

We think that extraction of domain knowledge in NLIDBs must not only reconstruct the information framework of the database and extract the semantic and usage information in the database schema, but also extract semantic and usage information based on user conceptual schema. Because most part of these information is not reflected in existing database schema and they are one part of user's own knowledge, it requires to be extracted by means of suitable learning mechanism.

## 2 Domain Knowledge's Classification and Extracting Strategies

Generally, the domain knowledge of NLIDBs can be classified into two types:

- *Domain-Independent Knowledge*, e.g. the usage of some common words, such as conjunction, preposition, all of which are independent on specific database application domain; and
- *Domain-Dependent Knowledge*, which is dependent on specific application domains.

Extraction of domain knowledge means to make up the gap between the realistic world knowledge model and database schema. Its goal is to provide necessary grammar and semantic information for the language processing in NChiq. Generally, domain-dependent knowledge falls into two catalogs:

1) Basic semantic information, which is contained in the database schema. The database logical model conveys some basic semantic information: entity and relation are expressed by tables, their names is expressed by symbol names; the primary key and foreign key describe the mutual reference relationship between entities, while relationship embodies what interplay relation of these entities and what positions these entities are; the constraint information depicts the restraint when an attribute is assigned a value.

2) Deductive semantic information, which is the extension to basic one and exists in domain application but is omitted when creates the database schema. Generally, such knowledge can be computed and deducted based on the basic information. For example, if the database stores the latitude and longitude information of different cities, then the distance between each city is the deductive information.

Because basic semantic information is dependent on relatively static database schema, the first phase of our extraction is DSE --- *Database Schema Extraction*. Generally, DSE can extract the following grammar and semantic information: entity, attribute, relationship (the verb usage of entity's names).

However, the design of database schema can not be controlled, some of which are non-normal (for example, they lack basic primary key constrain), so the database schemas maybe conform to different normalization. Because our extracting method is mainly based on heuristic rules, the above characteristic of non-normal gives rise to great difficulties.

Deductive semantic information exists beyond the database scheme of an application. For the complication of deductive information, the extracting of deductive information must employ different techniques to grasp the deductive

concepts as more as possible. We propose the following hybrid-extractor for deductive information:

- *Corpus based extracting*: acquire more grammar and semantic information, by collecting relevant query corpus and process of analyzing. For example, "live" is a verb that is not contained in basic semantic information, it is the verb connecting the entity "student" and the attribute "address", such potential word can be obtained only by plenty of corpus;

- *Data mining based extracting*: the database contains abundant semantic knowledge, which can be acquired by data mining.

- *Learn-by-query extracting*: while communicating with the user, the system learns new knowledge continually and enriches its domain knowledge.

### 3 Database Scheme Based Extracting in NChiql

NChiql is a Natural Chinese Query Language interface to database. In process of extraction, NChiql extracts the essential information of entities, relationships and attributes in databases by analyzing the relationship among these database objects. The procedure of DSE in NChiql is listed as following steps:

- (1) Get database scheme information through data dictionary;
- (2) Check whether every object is Entity or Relationship.
- (3) If it is Entity then extracting the semantics of entity;
- (4) If it is Relationship then extracting the semantics of relationship;
- (5) Extract the information of attributes of entity or relationship.

The extracting is mainly based on some heuristic rules. For example our judgement of entities and relationship is based on the following rules:

In a database schema  $R(R_1, R_2, \dots, R_n)$ , if in  $R_i$  there exists the primary key  $k_i (A_1, A_2, \dots, A_m)$ ,  $m > 0$ ,

- (1) If  $m=1$ ,  $R_i$  is an entity.
- (2) If  $m > 1$ , and primary key of  $R_j$  is  $k_j$ ,  $k_i \cap k_j \neq \emptyset$ , then  $R_i$  is relationship; otherwise  $R_i$  is an entity.

All the results of extracting are stored in dictionaries. NChiql divides dictionaries into two types, which are independent on domain and dependent on domain, namely general dictionary and specific dictionary.

- General dictionary is independent on application and is the core of the system, it records the semantic of words which are most commonly used, e.g. pronoun, conjunction, quantifier, interrogative, routine words, etc.

- Specific dictionary records the semantics and usage of words, which are usually used in a specific application domain. When the system is being transported to another domain, it must be reconstructed. The specific dictionary includes:

- (1) Dictionary of entity's semantics, which records the names of tables, synonyms and the semantics of the modifying attributes.
- (2) Dictionary of the attribute's semantics, which records the attribute's names, synonyms, hints, modifiers and the semantics of constraints.
- (3) Dictionary of the verb's semantics, which records the semantics and usage of the verb.

(4) Domain knowledge rules.

We call the results of DSE as *semantic conceptual model*, which reflects not only the semantics of words, but also, combination relationship among different words. So it's more powerful than the E-R model in term of semantics expression. The Figure 1 illustrates the conceptual model.

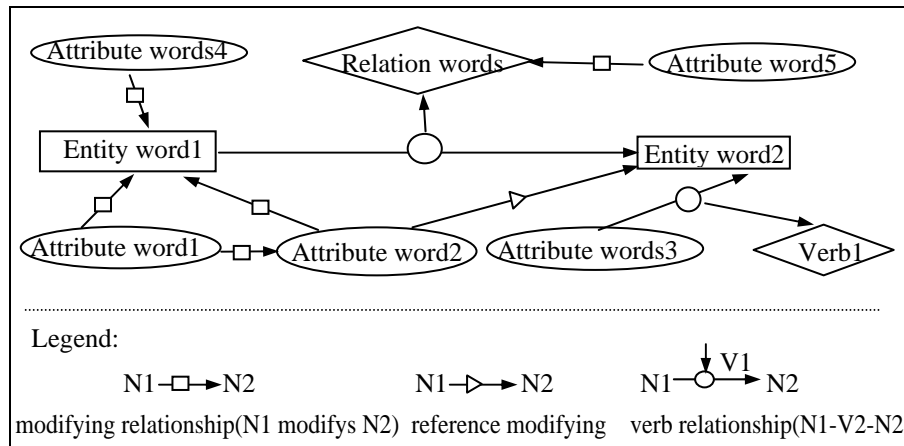


Fig. 1. Illustration of the semantic conceptual graph

The semantic conceptual graph regards semantic concept (words) as its main core, emphasizes on reflecting the semantic relationship among words. Generally, there are two types of modifying relationship and three types of verb relationship. To be specific, the types of modifying relationship are:

- (1) The Direct modifying relationship among entities, including the attribute-modify-entity and the entity-modify-entity.
- (2) Reference modifying among entities, which is engendered by foreign key reference relationship among entities.

The three types of verb relationship mean the verb combinations of relationships and entities, attributes and entities, entities and entities.

Modifying relationship can mainly determine the combination relationship of noun phrase(NP), while verb relationship presents the usage of verb phrase(VP). It is extremely useful to be provided with these semantic relationships in the analyzing process of natural language queries.

When NLIDB is transforming natural language into database language, it needs not only natural language knowledge, but also relevant database semantic knowledge, i.e. the mapping information from natural language to database semantic. NChiq1 is different from other extracting methods in that the usage of words is represented by the database semantic, namely the composition of words in a sentence and the relationship of each word are directly mapped to the table and column in database logical model. Such extracting method is word-driven and it is based on the semantic and functional characteristic of word.

In reality, because of the differences in the quality of the designers, there are so many database schemas that are difficult to suit the pre-assumption of DSE. To overcome these non-formal cases and alleviate the burden of the user as far as

possible, we propose some solutions, such as determining the primary key by reverse engineering [2], processing of non-normalization by building views, etc.

## 4 Related Works

TED[3] is one of the earliest prototype systems that dedicating to the problem of transportability of NLIDB. An important characteristic of TED is that it provided an automatic extracting interface--Automated Interface Expert, through which the user could guide the system to learn language and logical knowledge of the specific domain. The extracting ability of TED is very limited. Compared with TED, the extracting technology of TEAM [4] is relatively mature and complete. At first the extracting method of TEAM is menu driven and aided by interaction. In TEAM all the database objects to be extracted as well as the words extracted are listed in the menu, thus the whole extracting structure is very clear.

However, none of the above systems are able to take advantage of the information contained in the database schema sufficiently. Although TEAM includes extracting of the primary key and foreign key, it hasn't analyzed the relationship among entities and attributes, therefore it biased on extracting the general grammar of the words in a natural language category, thus increasing the complexity of extracting and leading to the complexity of language processing.

## 5 Conclusion

The semantic extracting is an indispensable part of natural language processing and the method of semantic extracting reflects the characteristic of the natural language processing. By extracting sufficiently the all sorts of information hidden in the domain, we could reduce the ambiguity of the words as far as possible and improve the efficacy of the natural language processing.

Now the difficult problem is the extracting of deductive semantic information. NChiqI has a perfect mechanism of extracting the basic semantic information, but lots of work should be done further on the extracting of deductive semantic information.

## References

1. Copestake, A., Jones, S. K.: Natural Language Interfaces to Databases. *The Knowledge Engineering Review* 4 (1990) 225-249.
2. Andersson, M.: Extracting an Entity Relationship Schema from a Relational Database through Reverse Engineering, <http://hypatia.dcs.qmw.ac.uk/SEL-HPC/Articles/DBArchive.html>
3. Hendrix, G.G., Lewis, W.H.: Transportable Natural Language Interface to Database. *American Journal of Computational Linguistic* 2 (1981) 159-165.
4. Grosz, B., et al.: Team: An Experiment in the Design of Transportable Natural-Language Interfaces, *Artificial Intelligence* 32 (1987) 173-243.