CLIR-Based Method for Attribute Correspondences Identification in Schema Integration

Hongding Wang[1,2], Yunhai Tong[1,2],
Shaohua Tan[1,2], Shiwei Tang[1,2], Dongqing Yang[1]

[1]School of Electronics Engineering and Computer Science,
[2]National Laboratory on Machine Perception
Peking University, 100871
Beijing, China
{hdwang, yhtong, tsh, tsw, dqyang}@db.pku.edu.cn

Abstract: In this paper, we discuss a new problem in heterogeneous databases, so called cross-language attribute correspondences identification, which is a more difficult problem than others in the field. Illumining by solutions of cross-language information retrieval (CLIR), we propose a CLIR-based method to deal with the problem. By studying the given example schemas, we analyze the problem in detail, and present the framework of a semiautomatic attribute correspondences identification system on basis of domain knowledge. We also give the identification procedure and discuss some problems relating the system, especially the difficulties building such a system. Moreover, we have developed a prototype to solve the problems of cross-language attribute correspondences identification we met in practice, which demonstrated effectiveness to identify cross-language attribute correspondences.

1 Introduction

Nowadays, databases interoperable ability becomes a crucial factor for the development of new information systems. Many diverse and diverging contributions have been made in the field of database integration [1-7]. Schema integration is the first step in database integration. There are three steps involved in developing an integrated schema [2]:

• Pre-integration, where input schemas are transformed to make them more homogenous (both syntactically and semantically).
• Correspondence identification, devoted to the identification and description of inter-schema relationships.
• Integration, to solve inter-schema conflicts and unify corresponding items into an integrated schema.

There are many problems when we integrate information sources. The usual one is naming conflicts, which arises when different names are employed to represent the same real-world fact (object, link or property). Naming conflicts are common in heterogeneous databases. The problem often occurs, especially in many large enterprises of China, because database systems were designed based on different

languages, for instance, with the same real-world concept, one system is in English language, while the other is in Chinese Phonetic Alphabet, i.e., PinYin. Therefore, the first step is to identify correspondences in schema integration, which is the foundation to deal with naming conflicts. Though attribute correspondences identification have been studied extensively in database integration, to our knowledge, we are not aware of any other work that considers identifying attribute correspondences involving cross-language. Facing such a new problem, how can we identify corresponding concepts? Methods of CLIR give us illumination to solve the problem [8-13].

The main intended contribution of this paper is to propose a CLIR-based method for dealing with attribute correspondences identification in schema integration, which relates at least two different languages. The rest of this paper is organized as follows. In Section 2, we review existing work in this area. Section 3 presents details of our method of attribute correspondence identification in cross-language heterogeneous databases, including some detailed discussions of such naming conflicts from databases, and how we use the CLIR-based method to identify attribute correspondences. Section 4 discusses the effectiveness of the method in application. Finally, we conclude the paper with a discussion of areas for future work.

2 Related Work

Both information retrieval and database integration are two important research fields in computer science. There are a lot of published technical papers on various aspects of the two problems in the ACM archive [1-13].

As is well known, the dual problem of synonymy and homonym is a major topic addressed by information retrieval research [11]. The dual problem exists not only in monolingual situation, but also in multilingual situation. Therefore, research in Cross-Language Information Retrieval (CLIR) attracts more and more interests of computer scientists. The research in CLIR explores techniques for retrieving documents in one language in response to queries in a different language [12, 13]. By either translating the queries into the language of the target documents or translating the documents into the language of the queries is the most obvious approach to CLIR. Translating documents is a very expensive task, most researchers in this field opted to take the query translation approach [12]. The query translation using bilingual dictionaries has been much studied by researchers in the field [8, 10]. Adriani et al. in [12,13] discussed some problems of dictionary based translation, and proposed some techniques to improve the effectiveness of the dictionary-based CLIR method.

In the research field of database integration, Larson et al. [14] discussed metadata characteristics and theory of attribute equivalence to schema integration. Based on types of metadata used, Li and Clifton [1] discussed three approaches for determining attribute correspondences: Comparing attribute names, comparing field specifications, and comparing attribute values and patterns. And they develop an automatic tool based on the metadata at the field specification level and data content level to identify attribute correspondence, which is called SEMINT. Moreover, Multi-User View Integration System (MUVIS) has addressed the problem of attribute correspondence, and the similarity and dissimilarity in MUVIS is primarily based on attribute names

[15]. Though comparing attribute names to identify correspondences has some limitations, yet it is a convenient and effective method to find correspondences in heterogeneous databases, because 'know the meaning directly from its name' is one basic principle during process of design database. Castano S et al. in [16, 17] made use of thesaurus to detect and solve naming conflicts, and they discussed three different kinds of thesauri to cover the terminology of the application domain. The first is a general, domain-independent thesaurus, which refers to the lexicon of a given language. An example is WordNet [18]. The second is a specific, domain-dependent thesaurus, which is constructed for terms used in the schemas. The third is a hybrid thesaurus, which can be constructed by adding specific terminological relationships holding for schema terms in the considered domain to enrich an available, domain-independent thesaurus for the domain.

In this paper, we mainly compare attribute names at the data dictionary level. We assume that attribute represented by synonyms in different databases is the same one. Therefore, if we find synonyms in different databases, we take for granted that they are attribute correspondences. As mentioned in section 1, attribute correspondences identification in multilingual databases is a new problem in database integration. Thus, we expect to deal with the problem by means of CLIR-based methods.

3 Cross-Language Attribute Correspondences Identification

We will illustrate the CLIR-based method dealing with cross-language attribute correspondences identification in this section. We start with two schemas from different databases of a large financial corporate in China.

3.1 Analysis of Cross-Language Schemas

Table 1 provides two schemas from different databases of a large financial corporate in China. According to the entity's name and attributes' name of schema 1, even there being abbreviations in the name of attributes, we can confirm that schema 1 must relate customer information, because it comprises much personal information, such as name, birthday and so on. In schema 2, we can not find any other useful information just looking at the schema, except for the data type of each attribute. However, those who are familiar with the database system from which schema 2 derives can easily find that schema 2 is similar with schema 1. How can they draw the conclusion? They depend on their work experience and domain knowledge. Usually, there is a manual named 'book of the database design in detail' before a database system construction, which will guide the database development and maintenance in the future.

Table 1 Sample schemas

SCHEMA 1	SCHEMA 2
CUSTOMER_INFOMATION (*CUSTOMER_NO: VARCHAR2(14),* *CUSTOMER_NAME: VARCHAR2(40),* *BIRTHDAY: DATE,* *SEX_TYPE: CHAR(1),* *CERTIFICATE_TYPE: CHAR(2),* *CERTIFICATE_NO: VARCHAR(20),* *NATION: VARCHAR2(30),* *EDUCATION_ID: CHAR(2),* *DEP_ID: CHAR(6),* *TENEMENT: VARCHAR2(128),* *ZIP_CODE: CHAR(6),* *PHONE_HOME: VARCHAR2(32),* *PHONE_OFFICE: VARCHAR2(32),* *EMAIL: VARCHAR2(60),* *COMPANY_TYPE: CHAR(2),* *OCCUPATION_NO: CHAR(2),* *COMPANY_ADDR: VARCHAR2(128),* *etc,* *)*	*KHXX(* *KHBH: CHAR(16),* *KHMC: VARCHAR(254,8),* *XB: CHAR,* *ZYDM: SMALLINT* *ZJHM: CHAR(20),* *JTZZ: CHAR(40),* *JTDH: CHAR(20),* *GZDW: CHAR(40),* *YZBM: CHAR(6),* *DWDZ: CHAR(40),* *DWDH: CHAR(20),* *JGBM: CHAR(9),* *ZJLX: CHAR,* *etc,* *)*

For example, figure 1 gives us examples of manuals' content of schema 1 and schema 2, respectively.

TABLE_NAME: CUSTOMER_INFORMATION

name in Chinese	*field name*	*data type*	*primary key*	*note*
客户编号 kè hù biān hào	CUSTOMER_NO	VARCHAR2(14)	TRUE	
客户姓名 kè hù xìng míng	CUSTOMER_NAME	VARCHAR2(40)	FALSE	
出生日期 chū shēng rì qī	BIRTHDAY	DATE	FALSE	
性别 xìng bié	SEX_TYPE	CHAR(1)	FALSE	m: male f: female

(a) A segment of schema 1

TABLE_NAME: KHXX

name in Chinese	*field name*	*data type*	*primary key*	*nullable*	*note*
客户编号 kè hù biān hào	KHBH	CHAR(16)	TRUE	NOT NULL	
客户姓名 kè hù xìng míng	KHXM	VARCHAR(254,8)	FALSE	NOT NULL	
性别 xìng bié	XB	CHAR	FALSE	NOT NULL	0: male 1: female

(b) A segment of schema 2

Figure 1 Segments of the database design in detail

From figure 1(we add the Chinese Phonetic Alphabet by full form PingYin under the Chinese phrases), we know reasons why both schema 1 and schema 2 represent

customer information. We also find naming specification of attribute in schema1 and schema 2: name of attribute in schema 1 bases on the English language, while name of attribute in schema 2 bases on the Chinese Phonetic Alphabet, using the initial of each Chinese character's PinYin, and their benchmark is the corresponding the name in Chinese, which represents the real-world concept. Those naming specifications are the domain knowledge. And such phenomena are common in database systems of large enterprises in China, especially in earlier systems. Here we call it multilingual problem in database systems. The problem arises from two reasons: a) there was no uniform specification in the large enterprises when they developed those database systems, b) English language was not as popular as today at that time in China.

Since we have unearthed secrets existing in schema 1 and schema 2, we can find attribute correspondence from those two cross-language databases. We bear in mind that the name in Chinese of attribute is the real-world concept in those two schemas, so the name in Chinese of attribute is a clue which helps us identify the attribute correspondences. For example, as figure 2 shows, we can find *CUSTOMER_NO* and *KHBH* represent the same real-world concept easily, so *CUSTOMER_NO* and *KHBH* correspond. In this paper, we denote attribute correspondences based on name as: $\Leftrightarrow_{name}$, so *CUSTOMER_NO* $\Leftrightarrow_{name}$ *KHBH*. Moreover, attribute correspondence based on name ($\Leftrightarrow_{name}$) has features as follow.

a) Reflexivity: $a \Leftrightarrow_{name} a$;

b) Symmetry: $a \Leftrightarrow_{name} b$, *then* $b \Leftrightarrow_{name} a$;

c) Transitivity: $a \Leftrightarrow_{name} b$, $b \Leftrightarrow_{name} c$, *then* $a \Leftrightarrow_{name} c$.

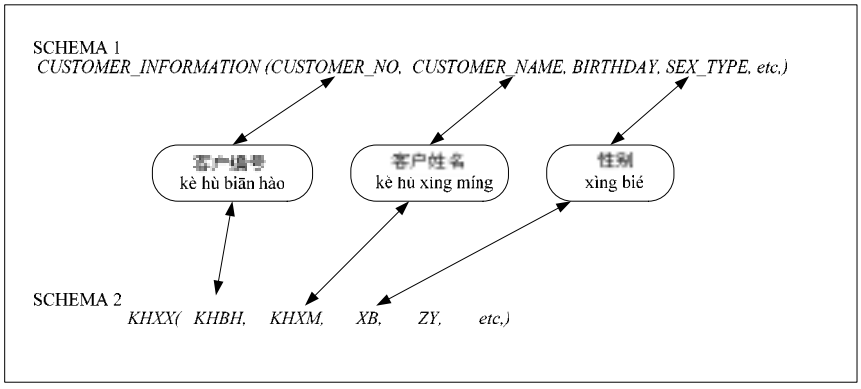

Figure 2 Example of attribute correspondences identification process

We also get that *CUSTOMER_NAME* $\Leftrightarrow_{name}$ *KHXM*, *SEX_TYPE* $\Leftrightarrow_{name}$ *XB* from figure 2.

3.2 Cross-Language Attribute Correspondences Identification Procedure

We have discovered rules between multilingual database systems, so we attempt to develop a computer-aided attribute correspondences identification system by means of methods of CLIR. We call it cross-language attribute correspondences identification system. An automatic or semiautomatic attribute correspondences procedure in such situation at least consists of English-Chinese translation, PinYin process and matching process, so the system comprises three main components: English processor, PinYin processor and matching module. Figure 3 shows the framework of the system.

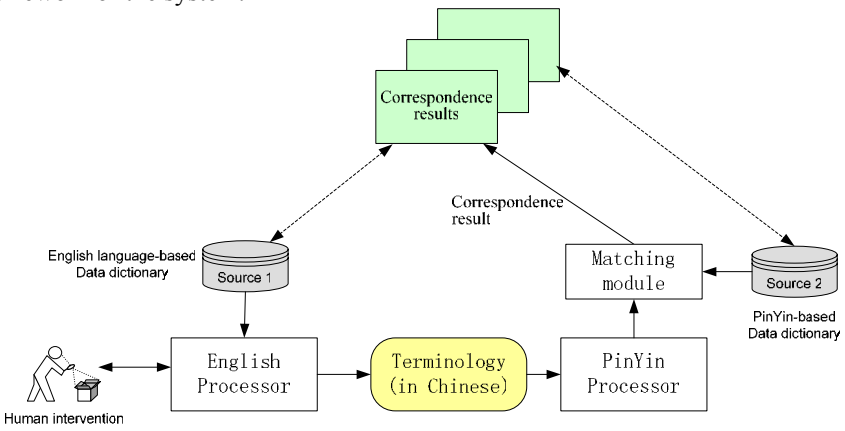

Figure 3 Framework of cross-language attribute correspondences identification system

We will discuss in detail the components of the system in the following paragraphs. As figure 4 shows, preprocessing component and English-Chinese translation component constitute English Processor. Preprocessing component mainly normalize

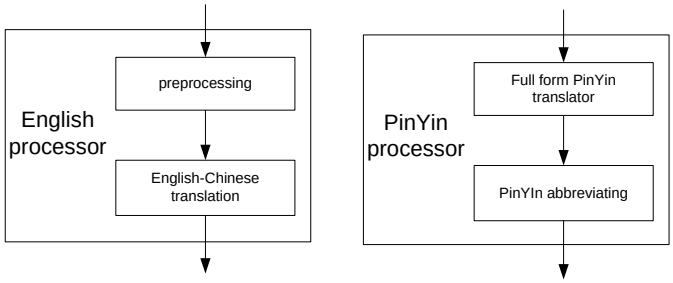

Figure 4 Components of English Processor and PinYin Processor

attribute names extracting from English language-based databases, for example, it restitutes abbreviations of attribute names to the full form (Number for NO, Employee for Emp, etc). After preprocessing, we get normalized attribute names of English language-based databases. Now we need English-Chinese translation component to translate attribute name into Chinese, which is the most difficult and onerous task in

the system. There are two ways to solve the problem. One is to make use of automatic or semiautomatic machine translation method [19, 20], the other is to use domain-dependent thesaurus [16, 17]. During English-Chinese translation procedure, human guidance and modification are necessary, which will be discussed in detail later. After translation, we get the terminology in Chinese, input of the PinYin processor.

PinYin processor has two components, Full form PinYin translator and PinYin abbreviating module. PinYin translator gives the full form PinYin of the terminology, which is not difficult assistant with existing tools and online Chinese dictionary. Based on the attribute naming specification in source 2, PinYin abbreviating module extracts the initial of the PinYin of each Chinese character, and forms a string according the order of the PinYin. The string is a candidate name of attribute correspondence.

On the basis of what English processor and PinYin processor having done, the Matching module works without too much difficulty. It extracts attribute names from PinYin-based database system, and compares those names with the strings from PinYin translator. If a name of attribute matches the string, the system finds a pair of attribute correspondences from the two heterogeneous databases.

Figure 5 illustrates diagrams of the attribute correspondences identification procedure in cross-language database systems. Of course, during identification procedure, PinYin abbreviating disposal is optional, because attribute name maybe use full form PinYin in some database systems.

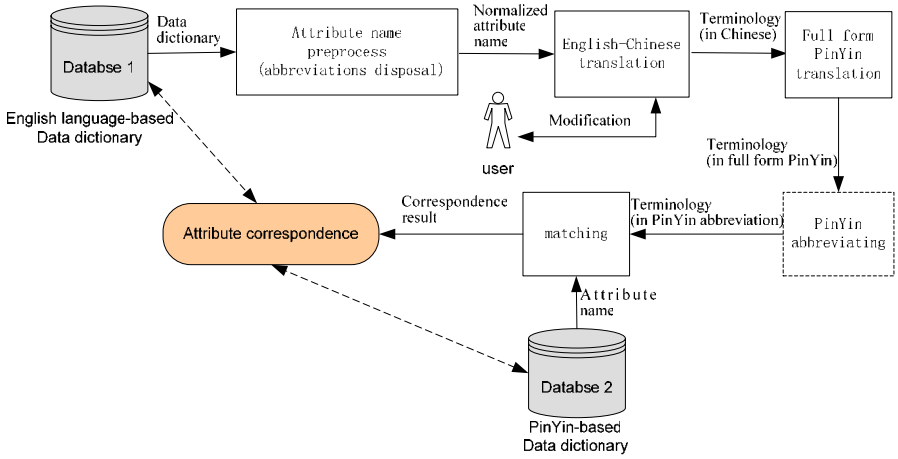

Figure 5 Attribute correspondences identification procedure in cross-language databases

Human guidance and modification are necessary in the procedure, especially during English-Chinese translating phase. Neither machine translation nor domain-dependent thesaurus works perfectly without human guidance. On one hand, machine translation usually produces term ambiguity [8-10], users should choose appropriate one corresponding the domain knowledge, on the other hand, construction a specific, domain-dependent thesaurus is an onerous task of users. Therefore, we suggest use both machine translation and domain-dependent thesaurus in the system. At the beginning, machine translation accomplish English-Chinese translation with user guidance and modification, and the system constructs a domain-dependent English-

Chinese thesaurus using the machine translation results at the same time. Along with the thesaurus becoming matured, the system can use the thesaurus doing English-Chinese translation. The system uses machine translation and updates the thesaurus when happens on strangeness. Of curse, the matured domain-dependent thesaurus can be reused in the same domain later.

4 Experiences with the Cross-Language Attribute Correspondences Identification Method

In this section, we will discuss our experiences of database integration practice in a large financial corporate of China using the cross-language attribute correspondences identification method. The corporate has a number of huge databases, and those databases were designed by different people for various purposes at that time. As a result, the format of data and its semantics presentation were not standardized. Now, the corporate begins to integrate these heterogeneous databases in order to centralize those data and construct an enterprise data warehouse. There are 17 databases involved in the enterprise-wide database integration. Table 2 generalizes some features of those heterogeneous databases.

Table 2 Some features of databases in a financial corporate of China

Database_ No.	Number of Tables	Number of Attributes	Naming language_Based	Note
1	37	362	PinYin	Initials of PinYin
2	211	3329	English	
3	238	5330	English	
4	124	1600	PinYin	Initials of PinYin
5	12	601	PinYin	Initials of PinYin
6	614	13000	English	
7	110	2403	PinYin	Initials of PinYin
8	44	551	English	
9	119	971	English	
10	74	722	English	
11	114	2278	English	
12	10	133	English	
13	63	1290	English	
14	51	563	English	
15	21	392	English	
16	63	406	PinYin	Initials of PinYin
17	180	2896	PinYin	Initials of PinYin
sum	2085	36872	PinYin:6 English:11	Number of attribute naming based on PinYin:8268

Table 2 shows that there are 6 database systems out of 17 assign name of their attribute based on the Chinese PinYin, which use initials of PinYin, the others based on English. As a matter of fact, there is plenty of overlapping information in those databases, such as customer information, product information and so on. Therefore, identifying semantically related objects and then resolving the schematic difference is the fundamental problem in those databases. Manually comparing all possible pairs of attributes is an unreasonably large task among those 17 database systems. In fact, we

take No. 6 database as the baseline system, because it is a core business system and will be de facto data specification according to the information planning of the corporate. So the reminder should be compared with the No. 6 database in order to identify the attribute correspondences. Our identifying practice shows that it requires an average of about 20 man-days per database to match elements when the task was performed by someone other than the data owner. If the name of attribute is not based on English, it would take more time. Moreover, there leave many mistakes in manual comparing result. It is no doubt that understanding the semantics of each individual database and integrating those heterogeneous databases are extremely difficult tasks in such complicated computing environments and with such large numbers of databases and attributes.

In order to reduce the work of people the problem, based on the thought of cross-language attribute correspondences identification method, we have implemented a prototype to accomplish attribute correspondences identification in such a situation. Though performance was not very perfect, the prototype has been effective for identification attribute correspondences in multilingual databases. With assistance of the tool, we saved about half time per database compared with manual identifying procedure, and reduced mistakes of attribute correspondences.

As we mentioned in section 3, human intervention is necessary during the identification procedure with the system. In fact, the most difficult and onerous work is English-Chinese translating phase, which takes almost 90% of manpower spent in the whole procedure with the tool. In the prototype, we use both machine translation and domain-dependent thesaurus to fulfill English-Chinese translation. When identifying correspondence between No.1 database and No. 6 database at the first time, we use machine translation directly to accomplish English-Chinese translation, and we modify the results of the translator, at the same time, we use the results constructing a specific, domain-dependent English-Chinese thesaurus. When the thesaurus becomes matured, the prototype can use the thesaurus doing translation, and human intervention will become less. Our experience shows that if the translation problem is solved successfully, the remaining work becomes easy.

5 Conclusions and Future Work

In this paper, we discussed a new problem of attribute correspondences identification in heterogeneous databases, whose names of attributes were named based on different languages. We analyzed the problem in detail, and presented a CLIR-based method for identifying cross-language attribute correspondences by means of domain knowledge and relationships of the databases. As far as we know, earlier work related attribute correspondences identification not mentioned this problem, and our results provided a first step in the direction.

This is a very preliminary work, and there are many issues to be studied in the future. Some examples are: to develop an efficient implementation of the system, to find self-learning methods to further reduce human intervention, to identify attribute correspondences together with other existing methods, and so on.

References

1. Li W, Clifton C., SemInt: a tool for identifying attribute correspondences in heterogeneous databases using neural network. Data Knowl Eng 33(1):49–84, 2000.
2. C. Parent, S. Spaccapietra, Database integration: An overview of issues and approaches, Communications of the ACM, 41(5):166-178, 1998.
3. J.M. Smith, P.A. Bernstein, U. Dayal, N. Goodman, T. Landers, T. Lin, E. Wang, Multibase integrating heterogeneous distributed database systems, in: Proceeding of the National Computer Conference, AFIPS, pp. 487-499, 1981.
4. Sheth A., Larson J. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys, 22(3): pp.183–236, 1990.
5. D. McLeod, D. Heimbigner, A federated architecture for database systems, Proceedings of the National Computer Conference, Anaheim, CA, AFIPS Press, Reston, pp. 283-289, May 1980.
6. S. Busse, R.-D. Kutsche, U. Leser, H. Weber, Federated Information Systems:concepts, terminology and architectures, Technical Report Nr. 99-9, TU Berlin, 1999.
7. Jose Samos, Felix Saltor, Jaume Sistac and Agusti Bardes, Database architecture for data warehousing: an evolutionary approach. In DEXA '98, pp.746-756, 1998.
8. Adriani, Mirna and Croft, W. Bruce, The Effectiveness of a Dictionary-Based Technique for Indonesian-English Cross-Language Text Retrieval. CIIR Technical Report IR-170, University of Massachusetts, Amherst, 1997.
9. Adriani, Mirna. Using Statistical Term Similarity for Sense Disambiguation in Cross-Language Information Retrieval. Information Retrieval, 2(1): 67-78, 2000
10. Ballesteros, L., and Croft, W. Bruce, Resolving Ambiguity for Cross-language Retrieval. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.64-71, 1998.
11. Charles H. Heenan, A Review of Academic Research on Information Retrieval. http://eil.stanford.edu/publications/charles_heenan/ Academic Info Retrieval Research. pdf, 2002.
12. Adriani and C.J. van Rijsbergen, Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. ECDL '99, LNCS 1696, pp. 311–322, 1999.
13. Adriani and C. J. van Rijsbergen, Phrase identification in cross-language information retrieval. In RIAO'2000 Content-Based Multimedia Information Access, volume 1, pp. 520–528, Paris, France, April 2000.
14. J.A. Larson, S.B. Navathe, R. Elmasri, A theory of attribute equivalence in database with application to schema integration, IEEE Trans. Software Eng. 15 (4): 449-463, 1989.
15. S. Hayne, S. Ram, Multi-user view integration system (MUVIS): An expert system for view integration, in: Proceedings of the Sixth International Conference on Data Engineering, IEEE Press, New York, pp. 402-409, February 1990.
16. Castano S, De AntonellisV, De Capitani diVemercati, Global viewing of heterogeneous data sources. IEEE Trans Data Knowl Eng 13(2):277–297, 2001
17. S. Bergamaschi, S. Castano, S. De Capitani di Vimercati, S. Montanari, and M. Vincini, An Intelligent Approach to Information Integration, Proc. Int'l Conf. Formal Ontology in Information Systems, June 1998.
18. G.A. Miller, WordNet: A Lexical Databases for English, Communications of the ACM, pp. 39-41, November 1995.
19. Chen Zhaoxiong, Gao Qingshi, English-Chinese machine translation system IMT/EC, Proceedings of the 12th conference on Computational linguistics , Budapest, Hungry, August 1988.
20. R. F. Simmons, Technologies for machine translation, FGCS, 2(2):83-94, 1986.