A Data Dictionary System Approach for Database Schema Translation

Shi-Ming Huang, Huei-Huang Chen, Shing-Han Li, and Joseph Fong*

Database/Knowledge Base Research Group
Department of Computer Science and Engineering
Tatung Institute of Technology, 40 Chungshan N. Road, 3rd sec.,
Taipei, 104, Taiwan, R.O.C.
Tel: (886)2 5925252-3291 Fax: (886)2 5925252-2288
E-Mail: smhuang@cse.ttit.edu.tw
* Department of Computer Science
City University of Hong Kong, Hong Kong

ABSTRACT

Database system reengineering technique has been used to resolve the problem of converting from existing out-of-data database systems to new technique database systems in order to reduce the new system implementation cost. The technique consists of three parts: schema translation, data conversion, and program translation. The schema translation requires the transferring of existing database schema into the new database schema with better semantics. Recapturing the semantics is a complicated and difficult work. During the database design phase, the semantics has been lost from the conceptual data model to the logical data model. It is difficult to recapture it. After recapturing the semantics from the original conceptual database schema, we need also to have a sophisticated knowledge base to store the knowledge. This paper describes a new schema translation system, which can recapture the missing/hidden semantics of a database. The kernel of the system is an Extended Entity Relationship (EER) Data Dictionary System (DDS), which can store all the semantics of the new database schema.

1. INTRODUCTION

Over the past decade, tremendous improvements have been made in Database Systems (DBSs). Such systems have evolved from Hierarchy and Network DBSs, through Relational DBSs, and we are now seeing the emergence of Object Oriented DBSs. Further more, the recent rapid growth in database technology has actually encouraged more installations of database management systems in different kinds of organizations. In addition to new database system installations, there is considerable interest in converting conventional file-oriented systems to database systems and upgrading out-of-date database systems to new technique database systems. The need to compete, to reduce costs and to standardize operational procedures makes conversions to a new technology a necessity for many organizations. The fact that many large companies still have a large number of sequential file systems or out-of-date database systems indicates a strong need to convert such systems to a new database system for better management.

Database system reengineering technique has been widely used to solve out this problem. It consists of three elements: schema translation, data conversion, and program translation [1]. The schema translation technique involves semantic re-construction and the mapping of the original schema into the new schema. The global database schema of heterogeneous database systems can be built by using such technique to capture the semantic meaning of each local database. The database conversion employs this technique to recover the semantic of original databases into new databases.

Recapturing semantics is a complicated and difficult work in the schema translation. Traditionally, a database administrator determines the database schema. The database administrator may not fully understand the user's view of the real world. Therefore, the data semantics may be lost in the system analysis phase. During the database design phase, the semantic meanings may be missing in the logical schema once the conceptual model has been mapped into logical model. It is difficult to recapture it. Much research has been done to solve this problem [2], [3], [4], [5], etc. Unfortunately, these works can not recapture the missing semantics automatically. They require extra information or knowledge from users to identify the missing semantics. Furthermore, only limited missing semantics can be recaptured through their works.

In recent years, data mining technique has been applied to this area. The techniques have increasingly been employed to enable knowledge to be discovered from the data held in the database. The knowledge discovery process can lead to suggested additions or alternations to the know structure of the database.

After recapturing the semantic of the original conceptual database schema during the schema translation, we need also to have a sophistic knowledge base that can store this knowledge. Unfortunately, much current research misses over looks this step. Most of the research use only a flat file or a simple table to store these knowledge [1]. These structures can not fully represent the recapturing knowledge and cause a distortion problem during the database schema translation.

A DDS has been used in the current database system to assistant system engineers during the design and maintenance phases. It contains the semantic of the current database schema. Many researchers have also suggested the DDS will be the future integration and resource repository center [6], [7].

The motivation for this research is to develop a database schema semantic resource center for database migration, integration, evaluation and conversion by using a data dictionary system approach. The kernel of the database semantic resource center will be the EER DDS. Fig. 1 shows

0-7803-4053-1/97/$10.00 © 1997 IEEE

the framework of the approach for database schema translation by using EER DDS.

The main focus of this paper is to investigate a default automatic schema translation mechanism for this research. The author implements a test-bed system, DARDSTS (Default Automatic Relational Database Schema Translation System), to prove our hypothesis. Due to the relational databases is most popular in the current information systems [1]. The EER model is commonly adopted in the database design phase and has more semantic than the relational model [1] [8]. Much research has also done to forward the EER model into relational model, object oriented model, etc [2] [9]. These two models have been chosen in the authors' test-bed system.

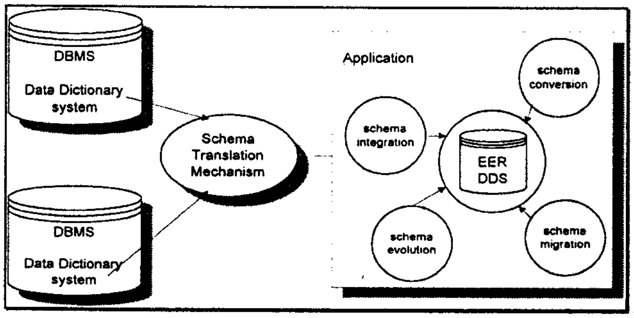

Fig.1. The framework of the approach for a database schema resource center

2. AN EER DDS

This section describes an EER DDS that is designed by the authors. In the authors' research, an EER model has been used to represent the original database semantic model after the schema translation. An EER DSS has been developed to store the semantic of the new database schema. Users allow to retrieve the semantic of the database through the system (see schema browser functions in section 4).

2.1 THE STRUCTURE OF EER DDS

The EER DDS is not like the existing DDS to support the full life cycle of an information system development, but to become the database schema resource center for the information systems reengineering [1]. There are four major design objectives for the authors' EER DDS:

- Contain the major features and capabilities found in the EER model
- Follow the ANSI IRDS (Information resource Dictionary System) structure.
- Be portable in RDBMSs (Relational Database Management Systems).
- Support a friendly user interface.

Fig.2 shows the meta-data of the authors' EER DDS. The meta-data is represented by EER Model and includes 10 meta-entities and 12 meta-relationships.

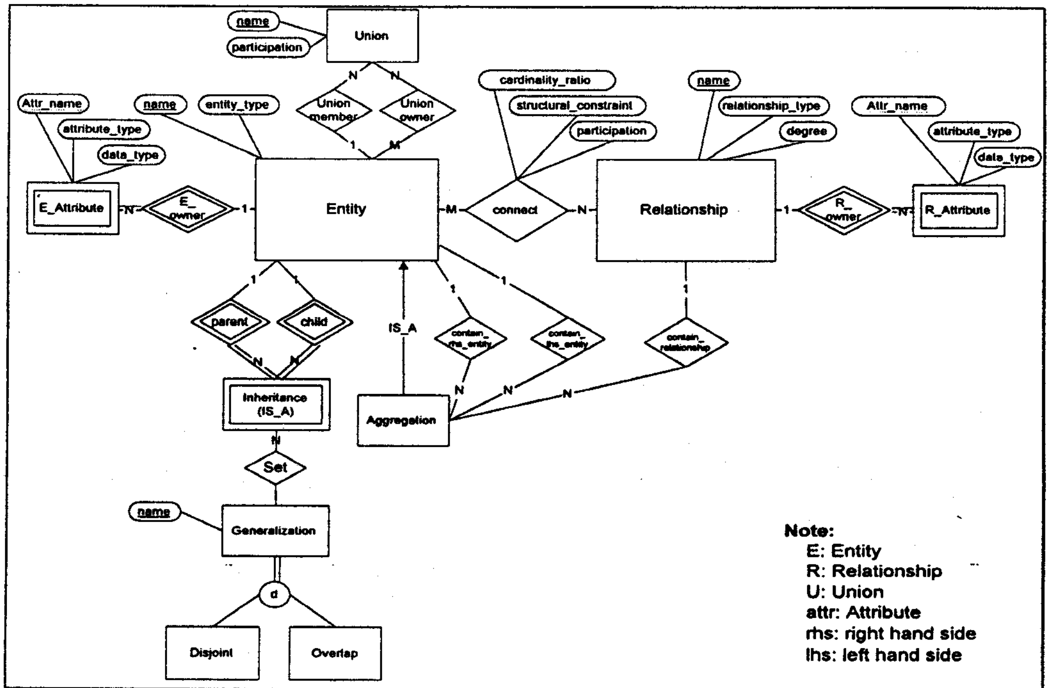

Fig.2. An EER database schema for the EER DDS

2.2 IMPLEMENTATION

Using the forward engineering technique [8], the EER database schema of the EER DDS can be mapped into the relational model to implement in a RDBMS. Fig.3 shows the EER DDS relational database schema that corresponds to the 4th normal form.

Entity(name,entity_type)
E_Attribute(attr_name,*owner_entity,attribute_type,data_type)
Relationship(name, relationship_type,degree)
R_Attribute(attr_name,*owner_relationship,attribute_type, data_type)
IS_A(*parent_entity,*child_entity,*generalization_name)
Aggregation(*name,*contain_rhs_entity,*contain_lhs_entity,

*contain_relationship)
Generalization(name, participation)
Disjoint(*name)
Overlap(*name)
Union(name,*member_entity, participation)
UnionOwner(*union_name,*owner_entity)
Connect(*E_name,*R_name,cardinality_ratio,
structural_constraint,participation)

note : The primary keys are underlined, and those prefixed with '*' are foreign keys.

Fig.3. An EER DDS relational database schema

The authors create a program module for the EER DDS. To enable the EER DDS can be installed at most of RDBMSs, the program module is coded by SQL.

3. THE DEFAULT AUTOMATIC SCHEMA TRANSLATION MECHANISM

This section describes the functions, which need for our knowledge discovery process to translate a relational database schema into an EER database schema as a kind of reverse engineering. The translation mechanism uses the systematic approach by accompanies with the data mining technique to recapture most of EER conceptual model semantics through the existing relational data dictionary system and the databases. Fig.4 shows the entire process.

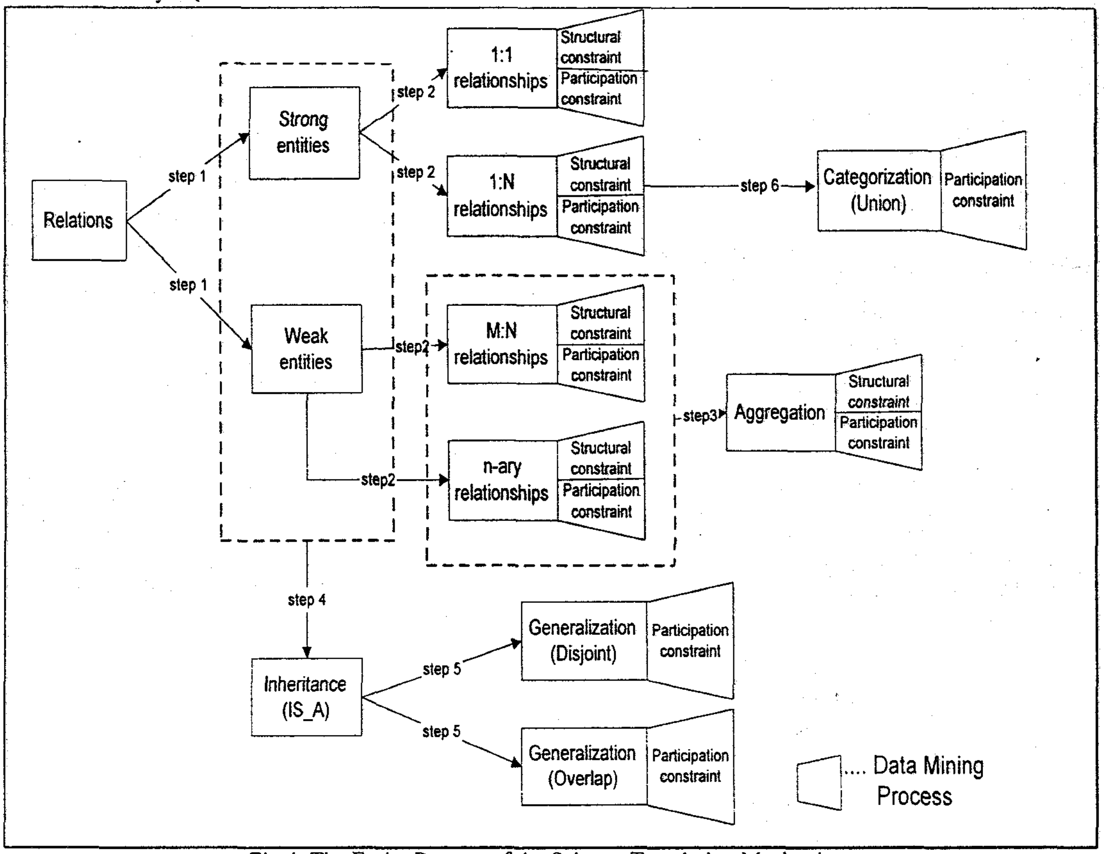

Fig.4. The Entire Process of the Schema Translation Mechanism

There are 6 steps within the translation mechanism. Each step represents different aspect to recapture the semantics of the database. The existing semantics of a relational database are retrieved from its data dictionary system. If any uncertainty or unsophisticated schematic is found, the data mining technique will be applied to discover more semantics from the real data. A reasonable result will be given as a default schema for the existing database after the data mining. The result only represents the situation of the current existing database. It stills is not the final database schema since some semantics has been missing. System designers must make the final decision for the database schema.

Pre-step: The pre-step is to construct the basic knowledge of the database for the schema translation. Firstly, the authors' EER DDS (see section 2) needs to be installed into the RDBMS, which stores the existing database. The EER DDS will become the database to store the halfway or final database schema during each step of the schema translation. Secondly, users need to identify the database, which will be translated. Thirdly, the entire entities of the existing database need to be recorded. Finally, the entire primary keys and foreign keys for each table of the existing database need to be found from its DDS. All these knowledge are stored into three temporary stand-alone tables.

Step 1. Classify entity types

The first step is to identify the entity type of a relation. There are two types of entities in EER model, i.e. a strong entity and a weak entity [10]. A relation whose identifier does not contain any foreign key is a strong entity; otherwise the relation is a weak entity. Fig.5 shows the process diagram of this step.

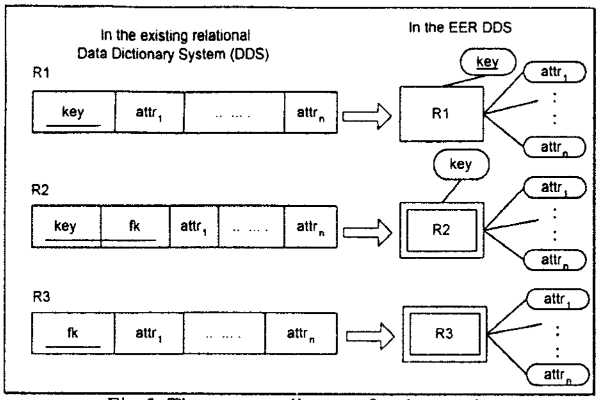

Fig.5. The process diagram for the step1.

Step 2 Discover binary and n-ary relationship semantics
This step is to recapture the cardinality semantics of relationships. The relationships are classified into two types: binary relationships and n-ary relationships. The cardinality ratio specifies the number of relationship instances in which an entity can participate. Common cardinality ratios for binary relationship types are one-to-one (1:1), one-to-many (1:N), and many-to-many (M:N). A relational model uses the foreign keys to represent the semantics of relationships and only supports 1:1 and 1:N relationship. To recapture the cardinality semantics of binary relationships, we can analyze the different types' entities by their foreign keys.

To discover higher level semantic for this issue, the data mining techniques are applied in the authors' research. The knowledge of participation constraint and structural constraint for the cardinality will be discovered after this step. The authors' mechanism can recapture these semantic from the existing DDS and store the new knowledge into the EER DDS. The following Fig. 6, 7, and 8 show the process diagram for the participation constraints.

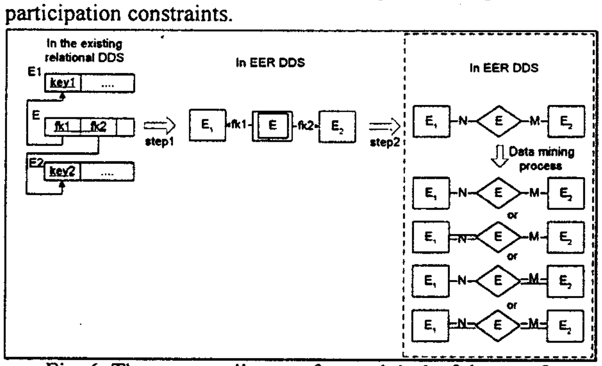

Fig. 6. The process diagram for module 1 of the step 2

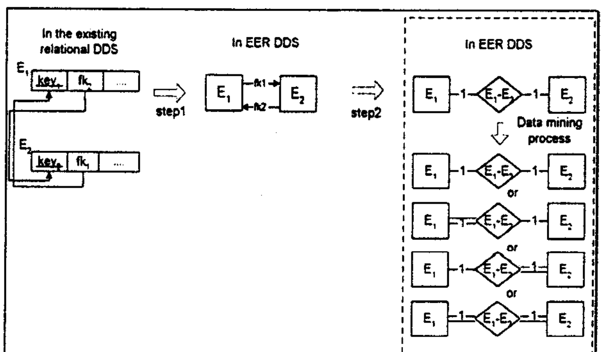

Fig.7. The process diagram for module 2 of the step 2

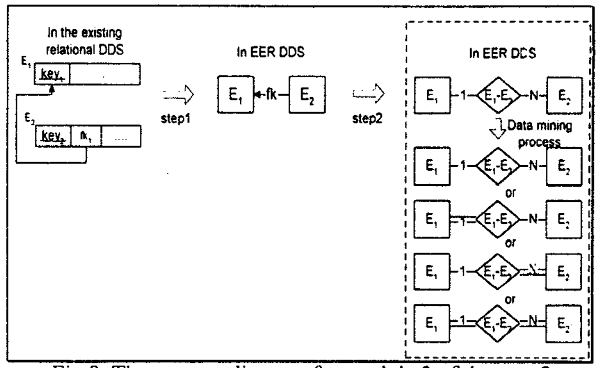

Fig.8. The process diagram for module 3 of the step 2

Step 3 Discover aggregation semantics
Aggregation is an abstraction concept for building composite objects from their component objects. Aggregation semantic can be represented by an aggregation entity. A binary relationship between two entities can be treated as an entity when the binary relationship relates to another entity. In relational model, this semantic can be represented in composite key without null value. Whenever a relationship relation is created, it must have a composite key pointing to the primary keys of two relations. This composite key appears in the third relation as a foreign key without null value. Fig.9 shows the process diagram of this step.

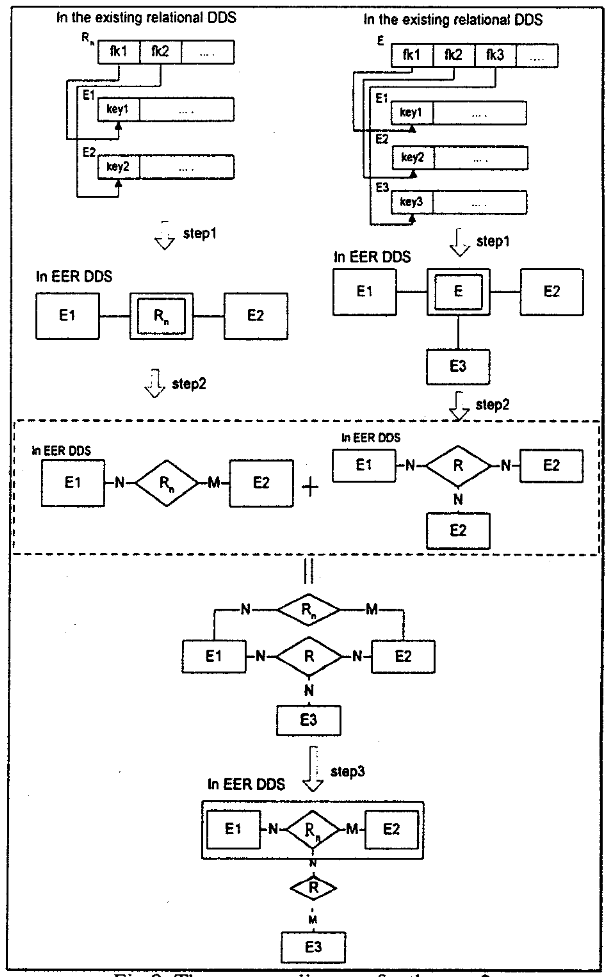

Fig.9. The process diagram for the step3

Step 4 Discover inheritance(is_a relationship) semantics

We can define an inheritance relationship as an entity being a member of a subclass, which inherits all the attributes of the entity as a member of the superclass. The entity also inherits all relationship instances for relationship types in which the superclass participates. Notice that a subclass, together with all the attributes it inherits from the superclass, is an entity type in its own right. The mechanism to recapture this knowledge is the following: if between any two or more relations there is an inheritance relationship then they have the same primary key and can distinguish the role of relation (i.e. parent or child) by their foreign keys. Fig.10 shows the process diagram for the step4.

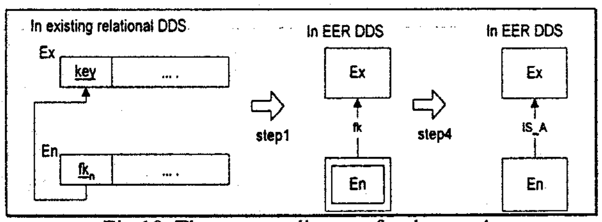

Fig.10. The process diagram for the step4.

Step 5 Discover generalization semantics

A generalization relationship is derived by mapping the "IS_A" relationship, which are connected by a common owner entity. Such semantics need to be confirmed by the analysis of the physical data in the database. There are two kinds of generalization: disjoint and overlap. The mechanism to recapture this knowledge is as follows:

- Disjoint -- A disjoint generalization is derived by mapping the "IS_A" relationship to owner entity and member entity in the EER data dictionary, such that the data of member entity are mutually exclusive to each other.
- Overlap -- Similarly, an overlap generalization is derived by mapping the "IS_A" relationship to the owner entity and the member entity in the EER data dictionary. However, there is no restriction on the data of the members.

Fig.11 shows the process diagram for the step.

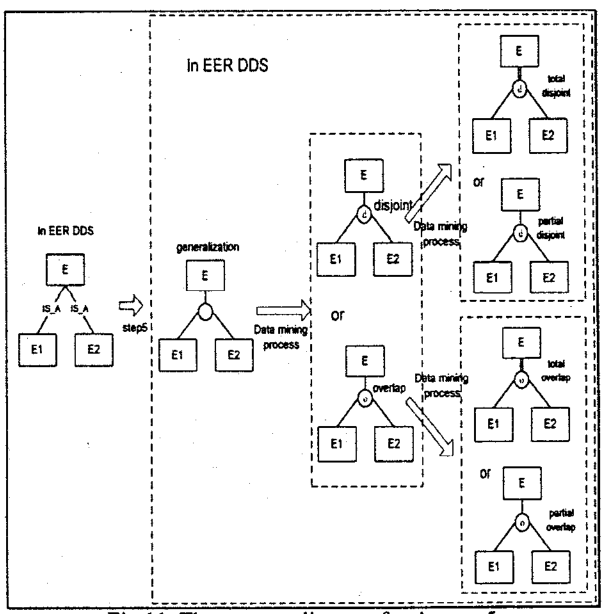

Fig.11. The process diagram for the step5

Step 6 Discover categorization (i.e. union) semantics

We have recaptured 1:N relationships in the step2. Several 1:N relationships may refer to the same entity and the primary key of these 1:N relationships is the same. This means that these 1:N relationships have a categorization relationship. This step is to recapture the categorization semantics from these 1:N relationships. Fig.12 shows the process diagram for this step7.

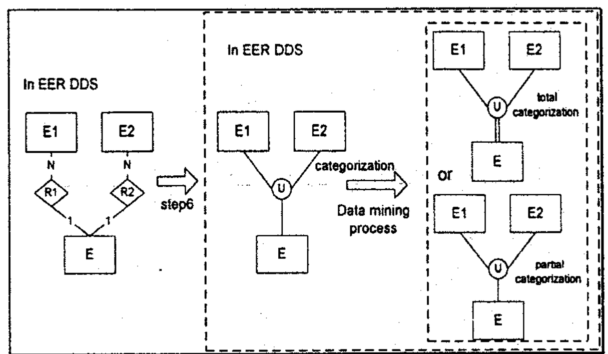

Fig.12. The process diagram for the step6

This is the final step for the authors' database schema translation. The results have show that more semantics have been found from the mechanism than relational database schema.

4. THE DARDSTS

A prototype system, called DARDSTS, has been designed by authors for this research (see Fig 13). The authors implement the DARDSTS on the client-server environment and use MS SQLServer and Gupt SQLWindows as the development tools.

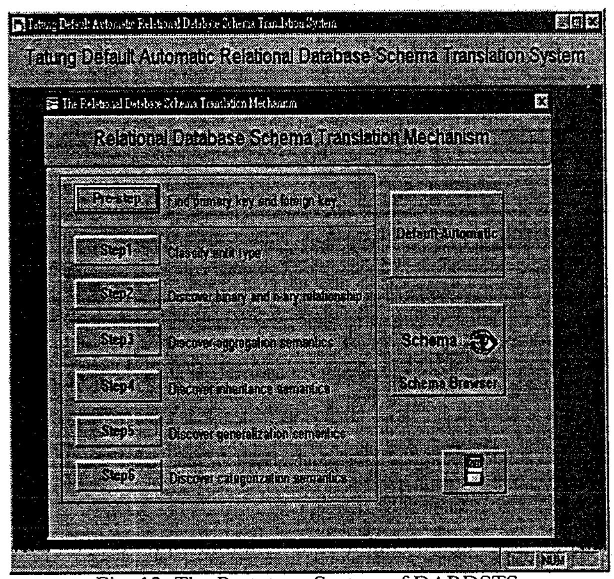

Fig. 13. The Prototype System of DARDSTS

The schema browser is the user interface of the EER DDS. It is a pull-down menu system (see Fig.14). The functions of the browser system consist of three parts: key, single table, and master-detail table. Users can use the push bottom icon to execute these commands.

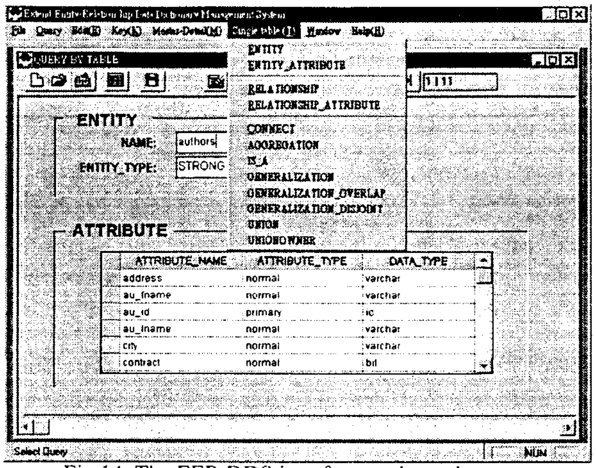

Fig.14. The EER DDS interface - schema browser

The commands under the key part include:

- Primary_key: List the entire primary keys about the tables, which are translated from the database.
- Foreign_key: List all the foreign keys about the tables, which are translated from the database.

The Commands under the single table part include:

- Entity: view the entities of the database by a table.
- Entity_Attribute: view the attributes of entities by a table.
- Relationship: view the relationship of the database by a table.
- Relationship_Attribute: view the attributes of relationships by a table.
- Connect: view the connect information between a relationship with an entity by a table.
- Aggregation: view the aggregation relationship by a table.
- IS_A: view the IS_A relationship by a table.
- Generalization: view the generalization relationship by a table.
- Generalization_overlap: view the overlap relationship by a table.
- Generalization_disjoint: view the disjoint relationship by a table.
- Union: view the union member by a table.
- Unionowner: view the union owner by a table.

The Commands under the Master-Detail table part include:

- Entity-attribute: using a master-detail table to display the attributes and their owner entity.
- Relationship-attribute: using a master-detail table to display the attributes and their owner relationship.
- Overlap: using a master-detail table to display the member entities and their owner entity.
- Disjoint: using a master-detail table to display the member entities and their owner entity.
- Union: using a master-detail table to display the owner entities and their member entity.

5. CONCLUSION

Database reengineering starts with the schema translation. Only after a schema has been translated can data and program then be reengineered into new database system. In this paper, the authors create an EER DDS, which is implement by SQL. The EER DDS can represent most of EER model semantics. This paper also has demonstrated a new mechanism to translate the relational database schema into an extended entity relationship model. The mechanism can default automatically recapture most of EER conceptual model semantic. The authors also implement this mew mechanism into system, called DARDSTS. The authors evaluate the system performance by using three different examples. These results also prove the success of this research [11].

Further more, the EER DDS can become the schema resource center, which can help the reengineer to do some applications, such as schema integration, schema conversion, schema migration, and schema evolution.

ACKNOWLEDGEMENTS

The work presented in this paper has been supported by the Research Community, Tatung Institute of Technology, Taiwan, under Grant No. B84016 and National Science Council, Taiwan, under Grant No. NSC 86-2213-E-036-005. We deeply appreciate for their finical support and encouragement.

REFERENCES

[1] Fong, J. and Huang S.M., "Information Systems ReEngineering", Springer-Verleg Publishing Company, Inc. ISBN 981-3083-15-8, 1997.

[2] Hughes John G. et al., "SYNDAMMA Methodology Description", SYNDAMA ESPRIT III-9006 report, 1996.

[3] Fong, J., Methodology for Schema Translation from Hierarchical or Network into Relational. Information and Software Technology Vol 34 No 3, March 1992, pp159-174, 1992.

[4] Put, F., "Schema Translation during Design and Integration of Database" Proceedings of 9th International Conference on Entity-Relationship Approach, pp399-421, 1990.

[5] Chiang H.L., Barron Terence M., and Storey Veda C., "Performance evaluation of reverse engineering relational databases into extended entity-relationship models". Lecture notes in computer science entity-relationship approach—ER'93, December 1993, pp352-363, 1993.

[6] Huang S.M., "An Information Resource Dictionary System for Expert System", Tatung Journal, Vol 25, Nov. 1995, pp281-291. 1995.

[7] Beynon-Davies, P., "Expert Database System - A Gentle Introduction", McGraw-Hill Book Company, London. ISBN 0-07-707240-5, 1991.

[8] Elmasri R., Navathe, S., Fundamentals of Database Systems(2nd edition). The Benjamin/Cummings Publishing Company, Inc. ISBN 0-8053-1748-1, 1994.

[9] Fong, J., "Mapping extended entity relationship model to object modeling technique, ACM SIGMOD RECORD, Vol. 24, No3., pp18-22, 1995.

[10] Batini, C., & Ceri, S., & Navathe, S., Conceptual Database Design: An Entity-Relationship Approach., The Benjamin/Cummings Publishing Company, Inc. ISBN 0-8053-0244-1, 1992.

[11] Li S. H., "Data mining relational database into extend entity-relationship model: a data dictionary approach," Tatung Institute of Technology, Msc. Thesis, June, 1997