SAMSTARplus: An Automatic Tool for Generating Multi-Dimensional Schemas from an Entity-Relationship Diagram

Jinho Kim[1], Donghoo Kim[1], Suan Lee[1], Yang-Sae Moon[1],
Il-Yeol Song[2], Ritu Khare[2], and Yuan An[2]

[1] Department of Computer Science, Kangwon National University,
192-1 Hyoja-dong, Chuncheon, Kangwon 200-701, KOREA
{ jhkim, korala765, webdizen, ysmoon}@kangwon.ac.kr
[2] College of Information Science and Technology, Drexel University,
Philadelphia, PA19104, U. S. A.
{song, rk84, yuan.an}@drexel.edu,

Abstract

This paper presents a tool that automatically generates multidimensional schemas for data warehouses from OLTP entity-relationship diagrams (ERDs). Based on user's input parameters, it generates star schemas, snowflake schemas, or a fact constellation schema by taking advantage of only structural information of input ERDs. Hence, *SAMSTARplus* can help users reduce efforts for designing data warehouses and aids decision making.

1 Introduction

Data warehouses maintain a collection of information which are extracted and calculated from online transaction processing (OLTP) databases. Most organizations are modeling their OLTP databases conceptually with Entity-Relationship Diagrams(ERDs). On the contrary, data warehouses are usually modeled as multidimensional structures (e.g., star schema, snowflake schema, constellation schema, etc.) to facilitate various online analytical processing (OLAP) operations [1]. To alleviate time-consuming design processes for data warehouses, several previous works developed manual or semi-automatic methods deriving multidimensional schemas from entity-relationship diagrams (ERDs) of OLTP databases [2,3,4]. We also developed an automatic tool, *SAMSTAR*, that generates star schemas from an OLTP ERD [5,6].

In this paper, we present *SAMSTARplus* that extends *SAMSTAR*. *SAMSTARplus* generates three different types of multidimensional schemas such as star schemas, snowflake schemas, and a fact constellation schema from an OLTP ERD according to users' options. The system will be able to help designers reduce the time for designing data warehouses and choose an appropriate multidimensional schema for the data warehouse.

2 The System Architecture

Fig. 1 shows the overall architecture of the *SAMSTARplus* system. The system takes as an input an ERD drawn by ERwin Data Modeler, stored in the XML format. The input ERD is modified by the Preprocessor into the one which is appropriate for the Schema Generator.

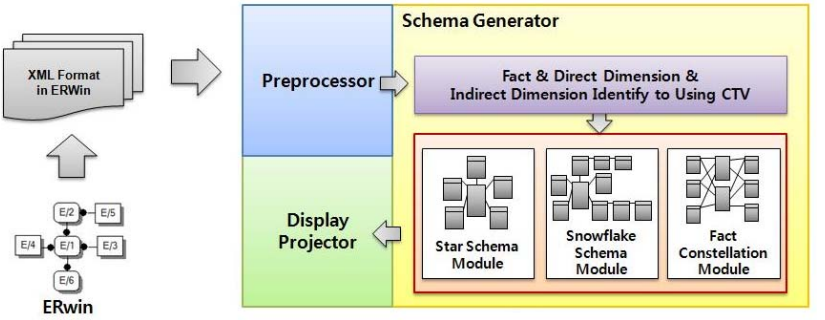

Fig. 1. The SAMSTAR*plus* system architecture.

The Schema Generator automatically generates star schemas, snowflake schemas, or a fact constellation schema from the input ERD. These three types of multidimensional schemas have a unique structural characteristic which each fact table is connected by several dimension tables. Thus, the fact table will come from an entity in the input ERD that is highly connected to other entities through M:1 relationships. We calculate this connection degree of each entity, called *Connection Topology Value (CTV)*, by the following equation [5]:

$$CTV(e) = w_d * count(Node(e)) + w_i * \sum CTV(Node(e)).$$

Here Node(e) represents entities directly connected to an entity *e*. CTV is calculated by both the number of these entities in Node(e) and the sum of their CTV values recursively (i.e., the number of entities indirectly connected to the entity *e*). w_d and w_i are weighting factors. For example, the entity E/3 of the input ERD in Fig. 2 (a) has one direct relationship with E/6 and one indirect relationship with E/7 (through E/6). Thus CTV(E/3) = 1*1 + 0.8*CTV(E/6) = 1*1 + 0.8*1 = 1.8 when w_d = 1 and w_i =0.8.

When an entity is related to more entities, it has a higher CTV value and has a higher probability of becoming a fact table. After calculating CTV for every entity in the input ERD, the Schema Generator selects as fact tables the entities whose CTV's are higher than a threshold value. (The threshold is an input parameter given by users as the condition for fact tables.) Fig.2 shows an example ERD and the CTV values of entities. If user's threshold is 5, E/1 and E/10 entities are chosen for fact table entities.

After selecting fact tables, the Schema Generator produces one of three multidimensional schemas (i.e., Star Schemas, Snowflake Schemas, and a Fact Constellation Schema) as the user wants. A star schema consists of a single fact table and several dimension tables. Thus, the Schema Generator produces one star schema for each fact entity (e.g., E/1 or E/10 in Fig.2) which has a dimension table for each entity (e.g., E/2, E/3, E/4, E/5, or E/6) connected

directly to the fact entity. When a directly-connected entity (e.g., E/2) has subsequent M:1 relationships (e.g, E/8), its dimension table is denormalized by combining all of them.

A snowflake schema is a refinement of a star schema that normalizes dimension tables. The Schema Generator builds one snowflake schema for each fact entity which has a dimension table for each directly-connected entity. The subsequent M:1 relationships of a directly-connected entity are maintained separately and referenced by the dimension table in the same way as in the input ERD.

A fact constellation schema has a complex structure that includes multiple fact tables simultaneously sharing dimension tables. The Schema Generator produces a fact constellation schema by a similar way as is done for a star schema.

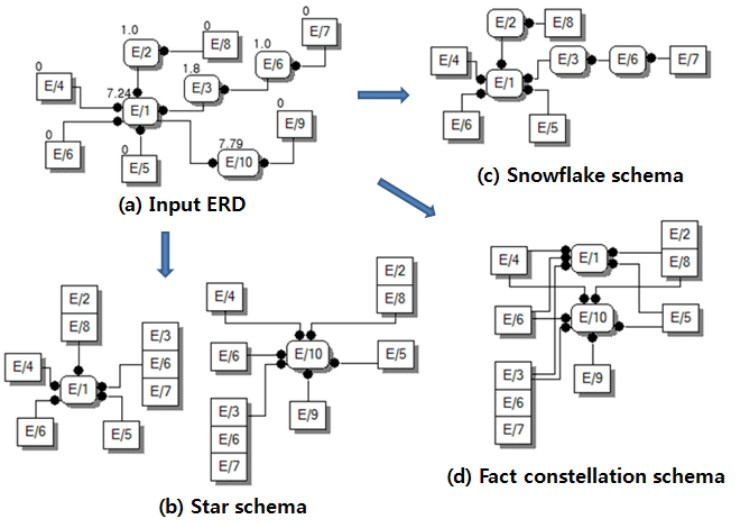

Fig. 2. An example ERD and multidimensional shemas

All of these multidimensional schemas are generated by taking advantage of only syntactic structural information of an ERD without using any semantic information or human interaction. Thus, the Schema Generator can produce them automatically.

The Display Projector presents graphically on the screen the resulting multidimensional schemas. It was implemented by using Jgraph, which is an open source graphic visualization library [7]. The JGraph was originated by the Swiss Federal Institute of Technology, and it can be used in any system of Java 1.4 or higher version. This Display Projector transforms the resulting schema into JGraph data structures interconnecting GraphCell objects within JGraph library. These JGraph data structures are finally displayed on the computer screen.

3 Demonstration Plan

This system runs at the site http://database.kangwon.ac.kr/samstarplus/. We demonstrate the system by applying several ERDs designed for example applications. The site provides example ERDs which are drawn by ERwin Data Modeler and stored in the XML format. By downloading them and specifying a threshold and a desired multidimensional schema type, users can get star schemas, snowflake schemas, or a fact constellation schema from an input ERD.

Furthermore, this demonstration will provide users with the chances drawing ERDs for their own applications by ERwin Data Modeler in on-site then generating their own multidimensional schemas through our system. This feature will be able to help users experience the advantages of our system.

Acknowledgement

This study was partially supported by the research grant from Kangwon National University.

4 References

[1] Chaudhuri, C. and Dayal, U. 1997. An Overview of Data Warehousing and OLAP Technology. In: ACM SIGMOD Record, Vol.26, No.1, pp.65-74, ACM Inc.

[2] Golfarelli, M., Maio, D., and Rizzi, S. 1998: Conceptual Design of Data Warehouses from E/R Schemes. In: 32nd Hawaii International Conference on System Sciences, pp. 334—343. Vol. VII, Kona, Hawaii.

[3] Husemann, B., Lechtenborger, J., and Vossen G. 2000: Conceptual Data Warehouse Design. In: 2nd International Workshop on Design and Management of Data Warehouses, p. 6, Stockholm , Sweden.

[4] Moody, D. L., Kortink, M. A. R. 2000: From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design. In: 2nd International Workshop on Design and Management of Data Warehouses, p. 5, Stockholm , Sweden.

[5] Song, I,-Y., Khare, R., and Dai, B. 2007. SAMSTAR: A Semi-Automated Lexical Method for Generating Star Schemas from an Entity-Relationship Diagram. In: 10th International Workshop on Data Warehousing and OLAP, pp.9-16, New York, USA.

[6] Song, I.-Y., Khare, R., An, Y., Lee, S., Kim, S.-P., Kim, J., and Moon, Y.-S. 2008. SAMSTAR: An Automatic Tool for Generating Star Schemas from an Entity-Relationship Diagram. In: 27th International Conference on Conceptual Modeling, pp. 522-523, Barcelona, Spain.

[7] JGraph site, http://www.jgraph.com.