Microsoft English Query 7.5
Automatic Extraction of Semantics from Relational Databases and OLAP Cubes

Adam Blum
Microsoft Corp, Redmond, WA. USA
mailto:AdamBlum@microsoft.com

1. What is English Query?

Microsoft English Query (EQ) lets users pose database queries in plain English. To do this, a developer need only define the database semantics, in effect building a conceptual model of the database. EQ provides an *Authoring Tool* that allows the developer to define the set of the entities and relationships in the database along with the database objects that the entities are associated with. Once this model is defined, the *English Query Engine* converts any English question posed in terms of the defined entities and relationships into a SQL statement. The application developer may use this engine, specifically it's a COM automation server, inside her Web, C++, Java, or Visual Basic application. This allows users to ask English questions of arbitrary complexity. These questions can be refined with as many follow-up questions as necessary) to find the information that is of interest.

EQ is a standard part of SQL Server 7.0 and 7.5. This talk explores the internals of a feature in the next release: the *Model Wizard*, which automatically extracts semantics from a database or cube and builds a model. It also describes *Semantic Modeling Format*, an XML grammar that allows any other tool to use the semantic information generated by the Model Wizard. Finally, the talk describes *Author by Example*, an amazing facility that, by analyzing failed questions, gleans additional database semantics that Model Wizard didn't capture.

Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999.

2.0. Model Wizard

The EQ Authoring Tool is easy to use and does not require any programming experience. The developer just needs to know the structure of their database. Still, it can be quite tedious to author semantic models for *very complex databases*. EQ 7.5, in beta test by VLDB99, contains a feature of particular interest to developers using very large relational databases or OLAP cubes. EQ uses a feature called the *Model Wizard* to automatically create the entities and relationships based on the database or cube structure. In effect, Model Wizard automatically extracts the database semantics from the database's tables and fields and the various objects in an OLAP cube: dimensions, levels, measures, properties and facts.

In order to capture as much of the semantics available in the database as possible, EQ has a rich set of heuristics for automatically determining the entities and relationships of a SQL database or OLAP cube. The talk describes these heuristics for both SQL databases, where they achieve approximately 70% capture of relationships that would be created manually, and OLAP cubes, where they automatically capture more than 85% of available relationships.

2.1. SQL Database Heuristics

The SQL database heuristics are divided into rules for creating entities and rules for creating relationships. Creating entities for each table and field is simple enough. Automatically creating relationships is more complex, and is driven by a set of rules, which include rules for creating:

trait phrasings between table and field entities
trait phrasings between table entities and other tables to which it have a join path..
name phrasings between certain field entities and major (table) entities
preposition phrasings between certain fields entities and the table entity

The talk describes these and other relationship creation rules in detail.

OLAP Cube Heuristics

The rules for determining entities and relationships in cubes are much richer. Thus they capture a larger fraction of the available relationships. The semantics of a cube are much more explicit than that of a database, due to the hierarchy of levels in a dimension, and the implied relationships of each of the dimensions and levels to the cube's fact and its various measures. Entities are automatically created for all OLAP dimensions, levels, properties (of levels), measures, and the cube's fact table. The OLAP relationship creation rules can be broken into the following categories:

level to level relationships – Each dimension contains a hierarchy of levels (if there is more than one level) that the Model Wizard exploits to create level to level relationships that reflect the hierarchy (lower level entities are "in" higher level entities which also "have" those lower level entities.) Also, due to the EQ engine's inference capabilities, the number of level to level relationships must be sufficient to be a "minimum spanning set" of the available levels with only one level of inference. This implies some interesting algorithmic challenges for level to level relationship generation that we will describe.

level to dimension relationships – The entity associated lowest level of a dimension is often the identifier or "name entity" of the dimension's entity.

level to fact relationships – Each level's entity is related directly to the entity that is associated with the cube's fact table.

dimension to fact relationships – Each dimension's entity is also related directly to the fact table.

fact to measures relationships – Each fact table "has" each measure that is available for the cube.

3.0. Semantic Modeling Format - Access to EQ Model Information

The Model Wizard Heuristics capture a wealth of semantic information about an OLAP cube or database. The authoring tool allows developers to add semantics that were not automatically captured. Since this information is a set of richly described entities and relationships, this conceptual model is of value for other applications.

EQ externalizes these extracted entities and relationships through an XML document following a grammar known as Semantic Modeling Format (SMF), described here and in the product documentation. SMF lets you either programmatically consume the EQ model or create it via the XML Document Object Model, a W3C standard API for manipulating XML hierarchies.

4.0. Author By Example

Any relationships not automatically created by the Model Wizard can usually be created in response to questions posed to the EQ model. This is possible through the Author By Example Feature. When working in the authoring tool, the developer can pose questions using a testing facility inside the tool. If EQ cannot answer the question directly, the developer can click the *Suggestion* button to invoke *Author By Example*, which will suggest a set of entities and relationships to be created which will allow the question to be answered. Author By Example can also be run in batch over a set of failed questions, for example captured from a Web site that incorporates English Query.

The talk will describe the architecture of Author By Example and how its designed to work to complement the semantic capture capabilities provided by the Model Wizard.

Each of these components will be demonstrated as part of the talk.