Advances in Engineering Research (AER), volume 105
3rd Annual International Conference on Mechanics and Mechanical Engineering (MME 2016)

Relational Database Ontology Discovery Method Based on Formal Concept Analysis

Zhi-Yong GAO[1], Yong-Quan LIANG[1,a] and Shu-Han QIAO[2]

[1]College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, P. R. China

[2] School of Forest, Shandong Agricultural University, Taian, P.R. China

lyq@sdust.edu.cn

Keyword: Database, Ontology, Formal concept analysis.

Abstract. This paper elaborates how to use the formal concept analysis method to map contents in the database to the ontology, in order to provide the big data application with high-quality data source by virtue of integrating the database with Semantic Web. In recent times, a mass of data is stored in the relational database, but such data with low share usage fails to play its full role. On account that the big data application has grown by leaps and bounds, a large number of shared data is urgently needed. By mapping the data in the relational database into the ontology, the technology of Semantic Web can provide a lot of semantic data to the big data application, which is conducive to big data analysis and use. In this paper, the ontology is built by taking the formal concept as the intermediate model and converting the logic structure of database into Hasse graph and context table, and then combining with the domain knowledge. The ontology in the knowledge domain can be found from the database by applying the formal concept analysis method, which takes full advantage of logical structure information of the database and is beneficial for automation found by the ontology. Eventually, ontology method and problems found in the relational database by virtue of the formal concept analysis are summarized herein..

Introduction

At present, the Internet develops at top speed; especially big data and mobile Internet technology grow up rapidly. Information resulted by people on the Internet grows at an exponential rate. The development of big data requires us to have more high-quality shared data. Moreover, if these data semantics are machine-readable, we can gain more accurate and valuable information from the data. The information sharing technology of Internet has developed from early HTML (plain text) to XML; from only providing text to providing pictures, sounds, video and other multi-media data; from unstructured data to semi-structured data. It can provide people with abundant multidimensional information, which brings about the revolutionary reform to people's work, learning and living, as well as the whole society. However, these technologies mostly serve people's reading and comprehension, which is not conducive to reading and analysing computer procedures. In order to directly exchange and share data between different computer systems, W3C (World Wide Web Consortium) proposes to serve XML as data exchange mark-up language first, and XML, based on meta-data, can customize the label. XML provides data between computer systems with a data type standard. With respect to these problems, Tim Berners-Lee formally put forward the concept of Semantic Web at the XML2000 meeting held in 2000, and also published the paper themed as "The Semantic Web" [1] on Scientific American in May 2001. Figure 1 shows the hierarchical structure of Semantic Web. W3C formulates the standards, such as RDF/RDFS(Resources Description Framework/Resources Description Frame work Schema) and OWL.

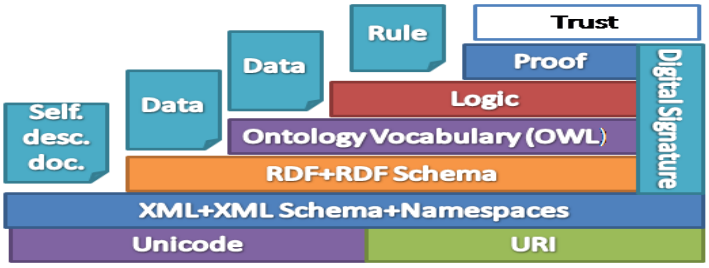

Figure 1. Hierarchical Structure of Semantic Web

In philosophy, ontology is the nature of the world. Although there is a variety of kaleidoscopic ontology's expressions and manifestations, the ontology is exclusive and remains unchanged. This is the theoretical basis in which the built ontology can map, integrate and share. In the Semantic Web, ontology is the core of Semantic Web, because it gives out the network resource's semantics. "An ontology is an explicit specification of a conceptualization" [2]. The ontology is a conceptual symbolic system, which is also a data model. Compared with other data models, such as Entity-Relation model, the ontology is rich in expression in the semantic concept hierarchy.

In the research field of Semantic Web, many researches do a good deal of work in mapping and transition between relational database and the ontology, as well as the technology of applying ontology to design the data base, and propose some semiautomatic and automatic methods. Mona Dadjoo divides the mapping method between the ontology and the database into three types: method based on logic model, method based on concept model and intermediate model method based on concept; furthermore, he puts forward an automatic mapping method, which also takes the trigger into account [3].

Relational Database

The relational model is proposed in the paper published by Ted Codd from IBM Institute in 1970. The model takes a mathematic relation concept similar to the value table as its fundamental part, and takes the set theory and the first-order predicate logic as its theoretical basis [8]. The relational model has developed rapidly upon the implementation of business in 1980s. The relational database system is generally applied for the storage of business system and website backstage data. Some entities are stored in the relational database.

Concept of Relational Model

The relational model expresses the database as a set of relations. When one relation is regarded as a value table, each row in the table represents a set of relevant data values. In the relational model, each row of the table indicates a fact generally corresponding to one entity or connection. Names of table and column are used for helping explain the meanings of every row of values. From the perspective of logic description, each row can be regarded as an assertion. Also, the relational model can be taken as Abox (Assertion Box) [4] for logic description.

The relation schema R may be defined and expressed by R(A1, A2,....,An) in this way. They consist of relation name R and the list of attributes A1, A2,..., An. Each attribute AI is the name of a role which is played by certain domain D in the relation schema. D is called the domain of AI, represented by Dom (AI).

For example, Table 1, in which every row represents a specific employee entity, is called Employee. Column name ..... designates how to explain the data values of each row based on the row where the value is. All values in the same column have the same data types. In the formal relational model terminology, the row is called tuple, the column called attribute, and the table called relation. The data type is called domain that describes the type of possible value in every column. It is a set of atomic values. As for the relational model, the meaning of atom refers to that every value in the domain is indivisible.

The relation can show two facts: one shows relevant entities, and the other one shows relevant connections. The relational model expresses the facts involving entities and connections as a relation. As a result, it is difficult to distinguish whether the relation shows the entity type or the connection. It must be distinguished according to the specific implementations. Another explanation of the relation refers to that the value in each tuple is explained as a value to meet the value of predicate. Such explanation is very useful in logic description, so the relational model is allowed to describe the logic.

Entity-Relation Model

While designing the information system, we build a data model for the real world. ER is the data model for entity relation. It believes that the world consists of entities and their relations. ER model is an important tool for database model. At the concept design phase of data base, the ER model and its varieties have been applied extensively. ER model is indicated by the ER drawing. Generally, the rectangle is used for indicating the entity type, ellipse for attribute, and rhombus or straight line for connection. The relation between the ER model and the relational model is shown in Table 1.

In practical use, the database design tool can be used to generate the ER model, such as SAP Sybase Power Designer, CA ER-Win, Microsoft Visio, Eclipse plug-in ERMater, and MySql Workbench.

Table 1. Relation between ER model and relation model [8]

ER	Relation
Entity Type	Entity relation
1:1 or 1:N connection type	Outer code (or connection relation)
M:N connection type	Connection relation and two outer codes
n-tuple connection type	Connection relation and n outer codes
Simple attribute	Attribute
Composite attribute	Integration of simple member attributes
Multi-value attribute	Relation and outer code
Value set	Domain
Code attribute	Main code (or auxiliary code)

In practical use, the database design tool can be used to generate the ER model, such as SAP Sybase Power Designer, CA ER-Win, Microsoft Visio, Eclipse plug-in ERMater, and MySql Workbench

Structured Query Language

Structured query language (SQL) is a comprehensive database language, including data definition, inquiry and sentence updating. Hence, SQL is a data define language (DDL), and also a data manipulation language (DML). Table, row and column in the SQL indicate relation, tuple and attribute in the relation data model respectively. On the basis of reverse engineering, it is easier to generate the logic layer of database from the database, i.e. DDL. The aforesaid ER model and EER model also need to generate the physical layer of database after converting into DDL, in order to build the storage structure of database. As regards the ontology mapping, the most important DDL sentence is CREATE TABLE. CREATE TABLE sentence includes all elements of the entity, which are the main contents for further ontology mapping. There is attribute data type and domain in the SQL. Basic data type available for attribute includes numerical value, character string, bit string, Boolean, date and time. The constraint plays a role in defining the relation between tables in

the mapping. But this definition in the concept hierarchy needs manually analysis, because it is not certain, rather than connotative.

Formal Concept Analysis

Formal concept analysis is an ordering relation based on mathematics, especially based on the lattice theory. It is a mathematical theory with regard to concept and concept hierarchy, which is put forward in the paper published by professor Wille, a German mathematician in 1982 [9]. In the formal concept analysis, concept, connotation and extension are defined strictly and clearly in mathematics. This also meets the requirements on clearing defining the concept. Moreover, the formal concept analysis may be served as the mathematical theoretical basis of the ontology. With the formal concept analysis, the concept hierarchy may be extracted from the dataset.

Formal concept and formal context are two basic concepts [10] of the formal concept analysis.

Definition 1. A formal context K:=(G,M,I) consists of two sets G and M, and the relation I between G and M, i.e. $I \in G \times M$. G element is called a formal object, and M element called formal attribute. In case of $g \in G$, $m \in M$, $(g, m) \in I$ represents that g has an attribute m.

Example 1 Articles in the office and their attributes are served as a formal context. It can be shown by one table, in which each row corresponds to one object, each column to one attribute, and the intersection of row x and column y is '×', representing that the object x has the attribute y.

Table 2. Context Table of Office Equipment

	a	**b**	**c**	**d**	**e**
1 Computer	×			×	×
2 Telephone	×	×			×
3 Table			×		×
4 Chair			×		×
5 Air conditioner	×				×
6 Wireless router	×	×			×
7 Sofa			×		×
8 Printer	×			×	×
9 Printing paper				×	

a Electricity utilization, b Communication, c Wood, d Word processing, e Fixed assets

Definition 2

Supposing $A \subseteq G$, we define

$f(A) := \{ m \in M \mid \forall g \in A, (g, m) \in I\}$ (a set of common attributes of objects in A).

Supposing $B \subseteq M$, we define

$f(B) := \{ g \in G \mid \forall m \in B, (g, m) \in I\}$ (a set of objects with all attributes in the B).

Definition 3. Formal context K:= two-tuple (A,B) on the (G,M,I), where $A \subseteq G, B \subseteq M$, and meets f(A)=B, g(B)=A. (A,B) is called a formal concept on the K, where A is an extension of the concept (A,B), while B is a connotation of concept (A,B).

Definition 4. If (A1,B1),(A2,B2) is two concepts on the context K:=(G,M,I), and $A1 \subseteq A2$, we called (A1,B1) the sub-concept of (A2,B2), where (A2,B2) is the hyper notion of the sub-concept of (A1,B1), and written for $(A1,B1) \leq (A2,B2)$, and the relation $\leq$ is called the hierarchical sequence of the concept. The set consisting of hierarchical sequences of all concepts of K is represented by B(G,M,I), called concept lattice on the context K.

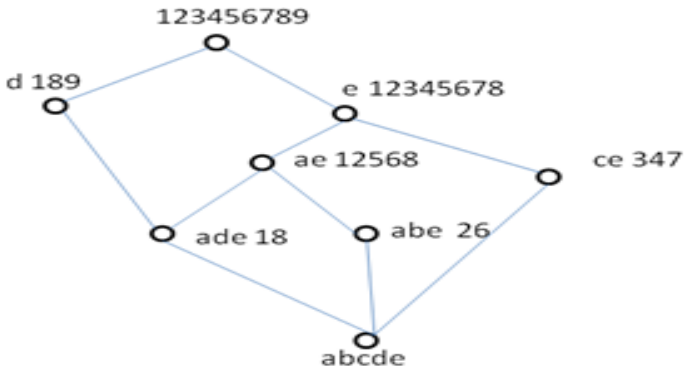

Figure 2. Concept Lattice of Office Equipment

In the real world, a variety of attributes of entity are not single-valued. For instance, the quality of articles may be divided into grades A, B and C based on advantages and disadvantages. Moreover, there is numerical value attribute of price. The concept scale is used for representing these attributes in the formal concept analysis. The basic idea is to convert the multi-value context into the single-value context by using the concept scale. The context with multi-value attribute is called multi-value context. After the Example 1 is simplified, the following table is obtained by adding two multi-values attributes (price and perfection degree).

Table 3. Concept Scale

	Electricity utilization	**Communication**	**Perfection degree**			**Price**					
			A	***B***	***C***	***≥ 10***	***≥ 100***	***≥ 500***	***≥ 1000***	***≥ 5000***	***≥ 10000***
1 Computer	×		×			×	×	×	×	×	
2 Telephone	×	×		×		×					
3 Table				×		×	×	×	×		
4 Chair				×		×	×	×			
5 Air conditioning	×				×	×	×	×	×	×	

See the References [6] [10] for details of formal concept analysis.

Ontology Language

A carrier is required for ontology. The carrier can be natural language, graphics, programming language, mathematics and other symbolic systems, or even a mixture of several symbol systems. In practical application, the researchers have developed a lot of languages used to describe ontology: RDF/RDFS, OWL, DAML-OIL and SHO. The important role of ontology is to realize the sharing of knowledge. An ontology language that can be widely recognized and used is what we need. At present, the OWL (Ontology Web Language) language standard developed by W3C in 2004 has been widely recognized and applied. OWL makes definitions for lexical semantics in RDF/RDFS on the basis of RDF/RDFS, facilitating the computer processing of meta-data in Web resources and realizing the interaction between machines.

RDF/RDFS

RDF is the standard recommended by W3C to describe resources, whose basic concepts are resources, attributes, and statements. The so-called resources can be all things that we can think of,

such as a piece of text, a number, an entity and a concept, which are not necessarily real ones. The characters in myths and legends can also be resources. Attribute is a special kind of resource, which describes the relationship between resources. RDF takes XML as its syntax. As XML has been widely used on the Internet, it is very conducive for the computer to automatically read or generate RDF. However, it cannot be considered that RDF and XML are the same; actually, RDF is a data model. RDF can provide an explicit expression of the relationship between entities, which is not supported by the XML document [7].

RDF uses URI (Uniform Resource Identifier) to identify resources, consisting of a unique resource identifier and an optional anchor ID. For example, a printer can be identified with "http://www.office.org/PID/20150781". It should be noted that this is not an accessible website, which is just similar in the form of the text.

URI is used to identify resources. Nevertheless, it is not enough just to identify resources; the ability of being able to describe the relationship between resources is also required, so as to establish the data model for describing the resources. It is achieved in RDF through statement. A statement is an assertion about a resource. The statement is a triple consisting of three parts: a subject, a predicate and an object. That is to say, RDF only provides binary predicate (attribute).

According to the RDF abstract data model, the order of descriptions (or resources) is inessential. Graph model is the real RDF data model. For example, the type representing a statement "http://www.office.org/PID/201

50781" is "HP7001". The data model of the statement is shown in the graph. The statement is shown as below with RDF:

Subject: http://www.office.org/PID/20150781

Predicate: http://www.office.org/Type

Object: HP7001

The object here can also be another resource, which is identified with URI.

RDF also provides a container element mechanism to organize resources.

rdf: Bag unordered container. There is no order between elements, and the repeat appearance of elements in the container is allowed, so it cannot be used to represent a collection.

rdf: Seq ordered container. There exists order between elements, and the elements can be appeared repeatedly.

rdf: Alt options collection. One is selected from multiple elements.

OWL

Although RDFS has extended the semantic representation of RDF, it still has a lot of semantic limitations, mainly including: non-intersection of defined classes, Boolean combinations of classes, cardinality constraints, and special properties of attributes (transitivity, uniqueness and inverse attribute). Therefore, W3C has developed Web Ontology Language (OWL) to further expand the concept semantic support. Ontology language is used to formally describe the display of the domain model. The main demands of ontology language are: syntax with good definition, efficient reasoning support, formal semantics, and full expression ability and expression convenience. Under the current technical conditions, it is unlikely to meet all demands of the ontology language. In order to meet the different demands of the specific applications, OWL defines three sub-languages: OWL Full, OWL-DL and OWL-Lite. OWL-Full uses all primitives, allowing the arbitrary combination of these primitives with the RDF/RDFS language, which is fully compatible with the RDF upward. For being able to perform efficient reasoning calculation, the sub-language OWL-DL Full limits the constructor use of OWL and RDF. On the basis of OWL-DL, OWL-Lite makes a further limitation that: owl:oneOf, owl: disjointWith, owl: unionOf, owl:complementOf, owl:hasValue and other constructors are not allowed to be used. Base statement can only take 0 or 1, which cannot be any nonnegative integer. Owl: equivalent Class statement can only be used for class identifier instead of anonymous class.

Ontology Discovery Method in Database

Relationship between Relational Database and Ontology

In practical applications, if someone wants to use ontology to share the contents of the database, there are two ways that can be adopted: (1) Logical concept model method. The data schema of the database is mapped to a subset of ontology, that is to say, the data schema is mapped to a subset of TBOX. The data in the database is considered as a subset asserted in ABOX, which is still stored in a database and used as an instance when it is required. The database is assessed to convert the data meeting the conditions to the required ontology instance. The method is suitable for large database, OLAP and data integration. (2) Database integral conversion method. The whole database is converted to a subset of ontology. Especially, the data in the whole database is fully converted to the ontology instance. This method is suitable for small database system and the condition that the data is relatively fixed.

Compared with ontology, the relational database lacks the relationship between concepts. Although the relationship between entities can be described with E-R graph, mainly including: 'is-a' and 'has-a', in the process of relational database design, the explicit distinction between 'is-a' and 'has-a' cannot be implemented in the database, not mentioning the hierarchical explicit expression between concepts. The hierarchical relations among relational database, FCA and ontology can be expressed by graph X: the relational database can be considered as a collection asserted in Object level and a subset of ABOX. FCA is the expression of Concept Level, which is mainly to express the hierarchical relationship between concepts. Ontology is used in Representation layer to describe the relationship between concepts. FCA is between ontology and relational database, effectively fitting the semantic difference [10] between relational database and ontology.

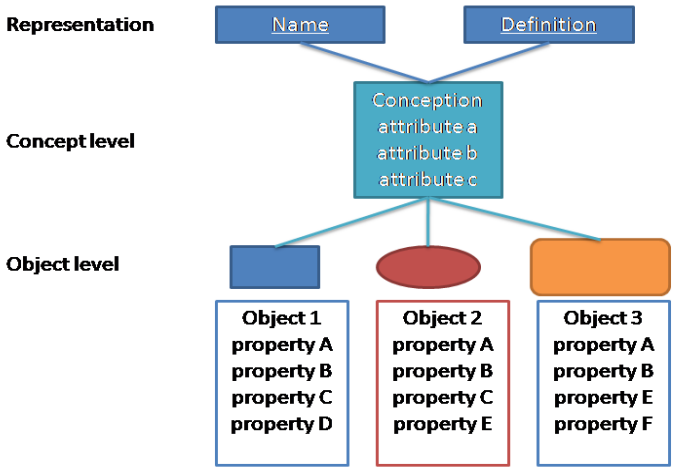

Figure 3. Structure of Concept Level

The differences between relational database and ontology are:

1)Unique name assumption. Relational database supports unique name assumption. For relational databases, the name of an instance is its unique identity, and instances with different names are different instances. It is usually PK of table in the concrete implementation. However, ontology does not support the unique name assumption. Even if the instances have different names, they also may be regarded as the same instances by inference engine.

2)The close world assumption and the OWA (open world assumption). The relational database is based on the close world assumption. It assumes that its own data and knowledge are complete. If someone queries a taxpayer's information in the database, and the database does not have information about the taxpayer, it will return 0/null, but ontology support is based on OWA, which

will give a response: no information available, when there is also no corresponding information for the problem mentioned above.

3)Theoretical basis differences. Relational database is theoretically based on relational algebra, with E-R used as a design tool. Ontology can be theoretically based on FCA, but it is not absolutely necessary. The semantic richness of ontology is much more than that of relational database [5].

Mapping Method between Database and Ontology

Great difference exists between relational database and ontology, which causes great difficulty to map the contents of the relational database to the domain ontology. However, the relevant research is rapidly developed for the mapping between ontology and relational database has huge application value. The mapping method between ontology and database can be divided into three types: method based on logic model, method based on concept model and intermediate model method based on concept. The method discussed herein is intermediate model method based on concept.

When implementing the mapping between database and ontology, a conceptual intermediate layer is added to solve the semantic difference between database and ontology. This difference is the main problem to be solved by mapping between database and ontology, that is, the main semantic richness of ontology is much more than that of relational database. Relational database does not explicitly express the concepts of the relationship between the entities. In this way, a conceptual intermediate layer can well fit the difference existing between ontology and database. At the same time, it is probable to verify the possible errors in the semantic logic of database.

Data Schema Acquisition

As for the mapping method between database and ontology, the first step is to acquire the data schema, which is mainly physical tables and views; a view is a logical table, depending on the other physical tables. Because of this, the view is an interpretation of the physical form from another logical point of view. The view provides more semantic information.

The data schema can be acquired from design document of the database, and a method of obtaining DDL text in the database by reverse engineering is also practicable. In practice, the method of using reverse engineering is more feasible because the database design document and the real database production environment usually tend to be different. The reverse engineering generating DDL can use various database management tools, such as management tools provided by Oracle, SQL server and MySQL and that provided by the third party database, which provide the function of DDL generation.

Formal Concept Analysis Based on DDL or E/R Model

The DDL text or E/R graph acquired is analysed to generate physical table and view table, which are converted to the context table of table XX. Then, Hasse graph of concept lattice is drawn based on the context table. What's more, they are stored as a table or graph database on the basis of the concept, attribute and relationship between them, which is required to be finished with the assistance of computer and human beings.

Conversion of Results of Formal Concept Analysis to OWL Ontology

The table or graph database is concerted to OWL ontology, which is automatically finished by computer, and then ontology is verified with inference engine. The concept, attribute and relation tables are converted to OWL ontology with ontology development tool Protégé [11] and ontology development kit based on Java. Online references will be linked to their original source, only if possible. To enable this linking extra care should be taken when preparing reference lists.

Conclusion

In this paper, the relational database mapping method is used to analyse the DDL or ER model of the database by FCA method, with the formal concept as the intermediate model which is converted to ontology expressed with OWL. The advantages of this method are that: (1) the formal concept is

used as intermediate model, to be able to clearly express the concept relationship in the database; (2) ontology in the database is easy to be found in an automatic manner; (3) formal concept analysis has a good foundation of mathematical theory [10], being conducive to the application of large data technology. Nevertheless, the research also has some problems, which need to be improved in the future researches: (1) the software tools used for formal concept analysis are insufficient, with heavy manual analysis workload; (2) the present method is mainly aimed at the logical structure of the relational database; further research is needed to analyse the content of the database.

Acknowledgment

Supported by the National Natural Science Foundation of China (Grant No.71403151).

References

1. Beners-Lee T., Hendler J. and Lassila O. The semantic web. Scientific American, 284(5): 34-43. (2001)

2. T. Gruber, A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition, , pp.199-220. (1993)

3. Mona Dadjoo,Esmaeil Kheirkhah. An Approach For Transforming of Relational Databases to OWL Ontology. International Journal of Web & Semantic Technology (IJWesT) Vol.6, No.1. (January 2015)

4. Anuradha Gali, Cindy X. Chen, et al., From Ontology to Relational Databases.International Journal of Web & Semantic Technology (IJWesT) Vol.6, No.1. (January 2015)

5. Michal Sir,Zdenek Bradac,Petr Fiedler,Ontology versus Database. IFAC-PapersOnLine 48-4 20–225. (2015)

6. Gerd Stumme, Formal Concept Analysis. Handbook on Ontologies 2ed. New York. 177-199. (2009)

7. Grigoris Antoniou,Frank van Harmelen, A Semantic Web Primer 2ed,The MIT Press Cambridge, London, (2008)

8. Ramez Elmasri, Shamkant B.Navathe,Fundamentals of Database Systems 6ed, Pearson Education, Boston,(2012)

9. R. Wille: Restructuring lattice theory: an approach based on hierarchies of concepts. In: I. Rival (ed.): Ordered sets. Reidel, 445–470. (Dordrecht-Boston 1982)

10. B. Ganter and R. Wille. Formal Concept Analysis: Mathematical Foundations. Springer, 1999. Translation of: Formale Begriffs analyse: Mathematische Grund- lagen. Springer, Heidelberg, (1996)

11. Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), (June 2015)