Knowledge Management in Construction - the framework of high value density knowledge discovery with graph database

JIANG Yong
School of Architecture, Tsinghua university, Beijing, China

WANG Ying-chu & WANG Zuo
UNIS Software Co. Ltd., Beijing, China

ABSTRACT: With the high speed development of information technology, the use of big data technology has been deepened into more and more areas. The outstanding problems of applying big data in construction industry are large scale, multi dimensions, strong supervision and serious information island phenomenon. How to obtain the useful business information (high value density knowledge) efficiently during the acquisition, storage and batch processing of large volume data is an important subject. Here we report a method of combining the graph algorithm and the traditional data warehouse concept to obtain high value density knowledge. This method uses the subjective judgment of human beings and the objective data analysis of the machines together to get more accurate results.

1 USING CONSTRUCTION INFORMATION CLASSIFICATION STANDARDS AND KNOWLEDGE TAG SYSTEM TO ENHANCE THE ABILITY OF KNOWLEDGE DISCOVERY IN THE CONSTRUCTION INDUSTRY YPE AREA

The construction industry is the pillar industry of the national economy and the first major industry. It covers all aspects of the building life cycle, closely related to a variety of urban construction projects. It also relates to the construction of the second industry and design, management, financial services and other related contents of the third industry. How to use big data, Internet of things, cloud computing, mobile Internet and other modern information technology to improve efficiency is an important subject relating to the development of the construction industry, and it is also a prerequisite link in the future construction of smart cities.

The data processing during the whole life cycle of a construction project is a typical nonstructural and big data use case, fully embodies the big data's 4V (Volume, Velocity, Variety and Value) character. The future development and intelligence of the construction industry is relying on the accumulation and continuing the refinement of the construction project data, digging out the relationship between the various data. So as to eliminate the information island phenomenon and break information asymmetry. According to the characteristics and contents of the construction industry, building a unified data platform (see Figure 1) based on the classification of construction industry information and knowledge management is the most basic and important step for interconnecting information and making full use of high value data. And the establishment of information classification method and keyword tag system is the key technology for this data platform.

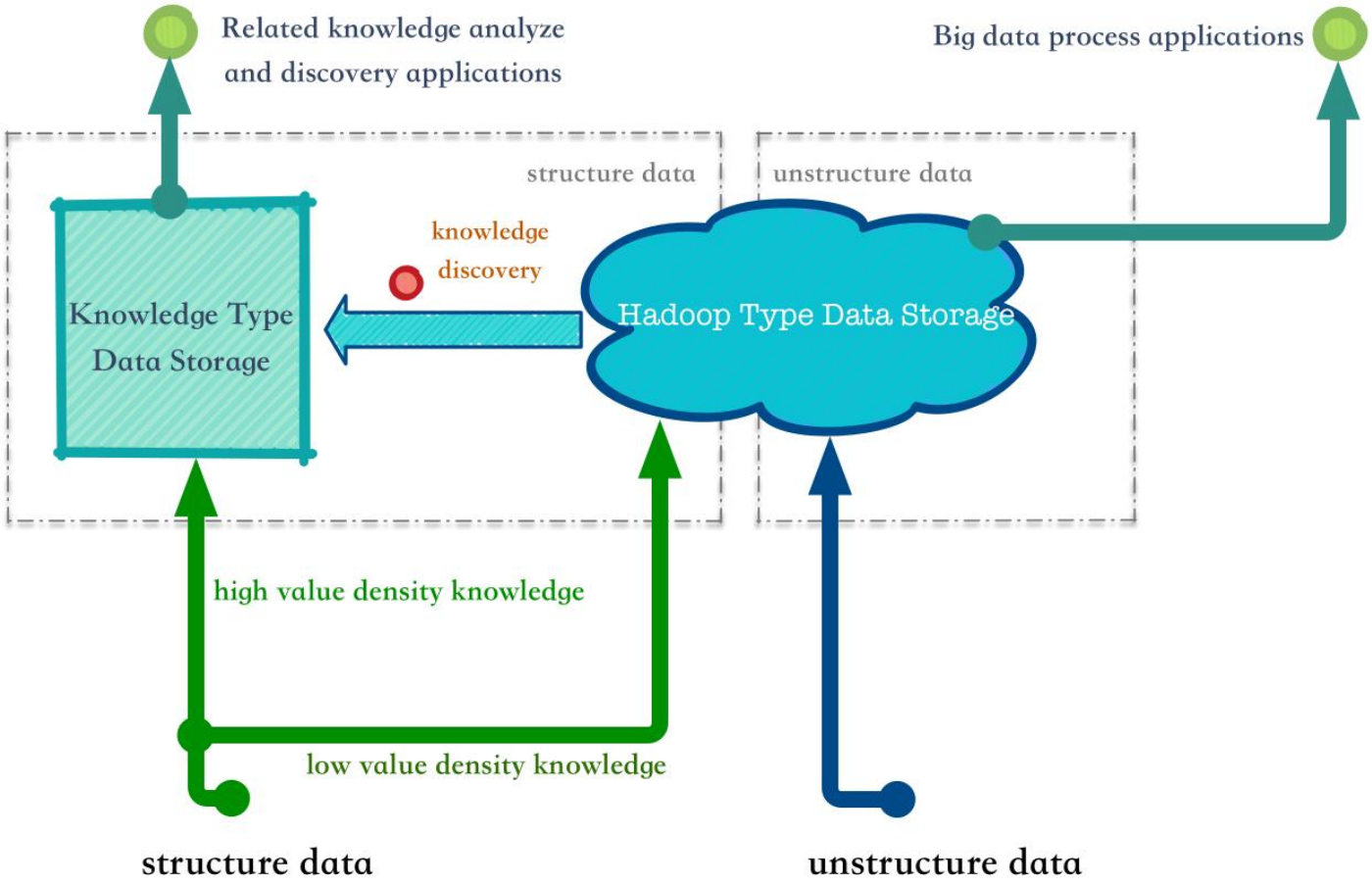

Figure 1: the data processing flow of the high value density knowledge discovery system.

The definition of information classification method and keyword tag system: A graph theoretic database structure based on IFC standard faceted classification. It stores all the information in the quality supervision platform of the construction project; uses data reduction and retrieves the relationship between different data on the subsequent development and evolution.

2 USING GRAPH DATABASE AS THE CORE INFRASTRUCTURE OF THE HIGH VALUE DENSITY KNOWLEDGE DISCOVERY SYSTEM IN CONSTRUCTION INDUSTRY

The main way to find high value density knowledge in the field of construction industry is analyzing, inquiring and exploring the relationship between different information classification tags attached on construction project data. Graph theory model is a priori advantage technology for handling this kind of "many to many to many" relation inquiries and the traverse computation as well as the quantitative analysis.

Graph theory is a branch of Mathematics, which is mainly about the study of graphs. The graph in graph theory is a shape which composes of a number of given points and lines (one line joints two different points), it is usually used to describe certain relations of some things, with points representing things, by connecting the two points lines represent relationship between two things. In the most common sense of the term, a graph is an ordered pair G = (V, E) comprising a set V of vertices or nodes or points together with a set E of edges or arcs or lines, which are 2-element subsets of V (i.e. an edge is associated with two vertices, and that association takes the form of the unordered pair comprising those two vertices). Usually, describing a graph is to draw points as a small circle, if the corresponding vertex has an edge, connect the two small circles with a line. How to draw these small circles and connections is not important, it is important to correctly reflect which vertices are connected by edges, which vertices are not connected by edges. Most areas in the real world can be modeled as graphs, such as social system, the recommended system and the association discovery system can be conveniently modeled as graphics. And the high value density knowledge discovery system discussed in this article is also a great scene for using graph theory model.

Graph databases are based on graph theory, they use graph theory to construct data access model, the whole data set in graph databases is modeled as a

large and dense network structure, so that graph databases allow simple and rapid retrieval of complex hierarchical structures. Most graph databases based on NoSQL structures such as key-value or document-oriented store. These storage engines have the concept of tags(properties) for data elements, it naturally fits the technic requirement of information classification method and keyword tag system and allows data elements to be categorized for easy retrieval.

3 DESIGNING HIGH VALUE DENSITY KNOWLEDGE DISCOVERY SYSTEM BASED ON GRAPH DATA MODEL AND TRADITIONAL DATA WAREHOUSE STAR SCHEMA CONCEPT

The value density of knowledge data isn't an objective value. It depends on the subjective needs of different users. In order to obtain the required high value density knowledge data, users must input a number of constraints (classification keyword tag information) as the basic filter for the selection of knowledge data. The more flexible users can set the constraint conditions; the higher accurate knowledge data users can get from system. On the other hand, the knowledge data stored in the system are combination of some scalar value, which is an objective fact. Since the knowledge data in system are limited objective value, and shared by all users, so that the constraints (classification keyword tag information) used for filter information should also be a set of describable measures within a limited scope. In order to implement these characteristics in graphic data model logically and systematically, we can learn the concept of star schema from the traditional data warehouse (based on the principle of relational database storage technology). The following is a brief definition of several concepts related to the traditional data warehouse and star schema:

Fact: The fact is the data unit in the data warehouse, and is also a unit in the multi-dimensional space, which is restricted by the analysis unit. Each fact includes the basic information about the facts (such as income, value, satisfaction, etc.), and is related to the dimension.

Dimension: The dimension is the space axis of the coordinate system. The coordinate system in the data warehouse defines the data unit, which contains the facts. In the data warehouse, the time is always one of the dimensions, other descriptive information with business implications are also main attributes used for defining dimension.

Star Schema: Star schema is a model that uses relational database to realize multidimensional analysis space (see Figure 2). It is a multidimensional data relationship, which is composed of a fact table (Fact Table) and a set of dimension table (Dimension Table). Each dimension table has a dimension as the primary key, all of these dimensions' primary key composite the primary key of fact table. The non-primary key attribute of the fact table is called Fact, which are generally numeric or other data that can be calculated, and dimensions are mostly text, time or other data types. By organizing data in this way, fact can be clustered and computed according to the different dimensions (part or all of the primary key of the fact table). By using this method, users can analyze the situation of business topics from different points of view.

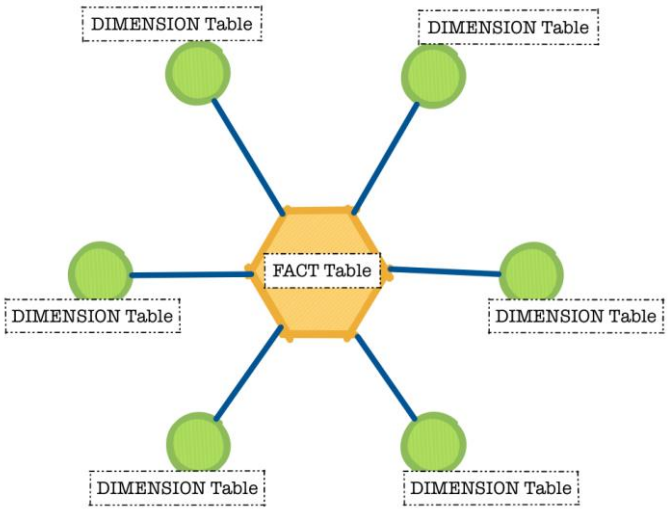

Figure 2: the internal database tables relationship of the star schema model.

On the macro, the internal components of star schema of the traditional data warehouse are essentially characterized by a graph. It just uses relational model in the specific data query and access operation. Inspired by this, we can use the graph data model to design an "improved" version of the star schema based on the graph database (see Figure 3) to serve the needs of the high value density knowledge discovery system.

The following is the main technical features of this new star schema model:

1. On design logic level, the model uses Fact to represent the concept of knowledge data and Dimension to represent the concept of constraint conditions.
2. On system implementation level, the model uses the vertex in the graph theory model to create the fact. The fact vertex is used to identify the objective scalar value of the knowledge data. These values are stored in the fact vertex as attributes.
3. On system implementation level, the model uses the vertex in the graph theory model to create the dimension. The dimension vertex is used to identify the constraint conditions users used to obtain high value density knowledge data. The dimension vertex only stores information associated

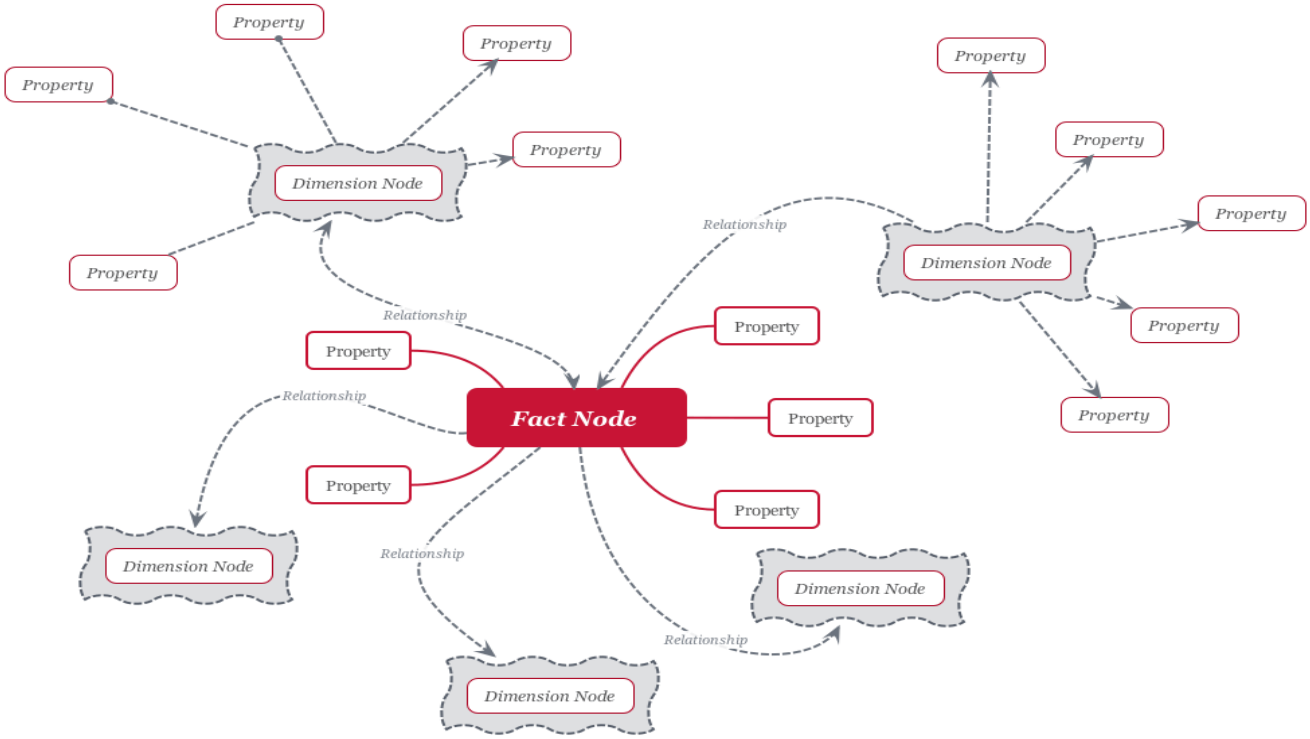

Figure 3: how to use graph data model to implement a "improved" star schema model.

with the constraint conditions. It does not store any data related to the knowledge data.

4. On system implementation level, the model uses the edge of graph theory to describe the relationship between Fact and Dimension. By setting the direction and different attributes of the edges, users can use the graph theory algorithm to screen the complex relationship between the fact vertices and the dimension vertices concisely.
5. Users can perform associated information discovery operations (multi node and multiple path correlation calculation algorithm in graph theory) on a particular dimension vertex to obtain the required high value density knowledge data (the fact vertices returned earlier by graph theoretic algorithms have higher value density).
6. Users can perform attributes query against fact vertices to select required knowledge data (similar to the traditional relational database data query).
7. When a specific fact vertex is obtained, according to their own subjective judgment, users can adjust and optimize the dimension vertices which are related to improve the accuracy of the value density of knowledge data under specific constraints.

4 CONCLUSIONS

The high value density knowledge discovery system which implemented on the framework we reported in this paper has the support of graph theory data model and algorithm. It can provide more powerful features and faster performance than the traditional relational model on discovering the association information under the multi dimension condition (the process of obtaining the high value density knowledge data). And by referencing the star schema concept in data warehouse field that has already been validated by IT industry, we can reduce the theoretical and practical risks of system design and implementation. At the same time, using the terminologies and operational methods that are familiar to the industry to obtain high density of knowledge data is conducive to the large-scale promotion and the use of the system.

REFERENCES

"From Relational to Graph Databases". Neo4j.

Ian Robinson, Jim Webber & Emil Eifrem. Graph Databases. O'Reilly, 2015.

J.A. Bondy & U.S.R. Murty. Graph Theory by Bondy and Murty. Springer, 2008.

Jiawei Han, Micheline Kamber & Jian Pei. Data Mining Concepts and Techniques. Morgan Kaufmann, 2011.

Len Silverston, W. H. Inmon & Kent Graziano. The Data Model Resource Book: A Library of Logical Data Models and Data Warehouse Designs. Wiley, 1997.

Richard J. Trudeau. Introduction to Graph Theory (Dover Books on Mathematics). Dover Publications, 2013

W H.Inmon. Building the Data Warehouse. Wiley, 2005.