Aggregation Strategy Based on Conversion from Dimension Classes to Dimension Hierarchies

Zhang Yan, Luo Xu

Teaching Department of Computer and mathematical foundation of Shenyang, Normal University, Shenyang, 110034, China.
Zhangyan_synu@126.com, Luoxu_2002@sina.com

Abstract - After analyzing the signification of structure and definition about dimension to aggregating operation, confirms that some dimension classes maybe produce analysis demands which are similar to dimension hierarchies, provides CDCDH method for conversion from dimension classes to dimension hierarchies. Gives definition on dimension class and dimension hierarchy, furthermore, gives definition on visible dimension hierarchy and hide dimension hierarchy. About hide dimension hierarchy which shows dimension class, confirms the condition and steps for conversion to apparent dimension hierarchy. In the meantime, produces the realizing procedure about CDCDH method through a practical instance.

Index Terms - Data Warehouse, Dimension Class, Dimension Hierarchy, On-Line Analytical Processing.

I . Introduction

To get a better application in aggregation operations on the dimension hierarchy, a strict dimensional hierarchy definition is required to implement aggregation operations effectively and accurately in the design of systems data warehouse functions systems. Literature [1] gives a short description of the dimension hierarchy definition as well as analysis operations of dimension. But reference [1] did not give the strict definition of the dimension hierarchy, and did not discuss dimensional analysis in depth, and a further study of the relationship between them was not given.

References [2] and [3] introduce the idea of saving running time by implementing the pre-determined aggregation computations on data cube with the result storing in a multi-dimensional storage. References [4] and [5] improve the algorithm on aggregation operations and its analysis. But, aggregation operations are based on dimensional structure definitions, OLAP does the analysis on dimension hierarchy and computing paths. Therefore algorithms on aggregation computation are required to be designed on dimension hierarchy. The mentioned articles focused only on the basic concepts of dimension hierarchy without exploring the generation and implementation of special dimension hierarchy. Reference [6] introduces the demand for the complexity analysis of dimension hierarchy from implementing the data warehouse. But Reference [6] gives no further research.

Due to the lack of in-depth study of the concept and mechanism of the dimension hierarchy, during the design of data warehouse, the design of dimension hierarchy with apparent basic aggregation operations is valued, but the design of hidden dimension hierarchy is ignored, it leads to the lower of OLAP database availability. Studies have found that hidden dimension hierarchy does exist and they are showed in dimension classes (classification of dimension), which implies that the dimension class requires similar operations to dimension hierarchy operations. So this paper expands the definition of dimension hierarchy, gives the concepts of apparent dimension hierarchy and hidden dimension hierarchy, and expatiates on the algorithm CDCDH (Conversion from Dimension Class to Dimension Hierarchy), which provides an effective solution to the generation of hidden dimension hierarchy.

II . The description of dimension classes and levels

Reference [1] pointed out that the multi-dimensional analysis based on dimension classes and dimension hierarchies are not the same. The former mainly means classification and induction. The latter refers to the process of generating data from bottom to top (Roll-Up) and opposite process (Drill-Down). In order to better discuss CDCDH method, this paper defines the dimension class, the dimension hierarchy, and further introduces the apparent and hidden dimension hierarchy concepts.

A . The Dimension Class

Set R is a binary relation defined on all tuples related to an attribute of the dimension table A, If R satisfies:

1). reflexivity, any a $\in$ A, then a R a;

2). symmetry, any a, b $\in$ A, then a R b => b R a;

3). transitivity, any a, b, c $\in$ A, the a R b and b R C;

Then R is called an equivalence relation on the dimension table. For example, in the teacher dimension table of the data warehouse, the relation on "Title" property with identical tuple values is an equivalence relation.

If R is an equivalence relation on the dimension table A, any dimension table tuple a $\in$ A, then the set of tuples $[a]_R=\{x|xRa, x\in A\}$ is the equivalent of a R class. The teacher dimension table of the data warehouse, in terms of equivalence relations on "Title" property, professors, associate professors, lecturers, teaching assistants is different equivalence classes.

For non-empty-dimensional table A, A_α is a nonempty sub set of A, where$\alpha\in$ I, I is a subscript collection of A_α, if it satisfies:

$$\bigcup_{\alpha\in I} A_\alpha = A, \text{ if } \alpha,\beta\in I \text{ and} \alpha\neq\beta, \text{ then } A_\alpha \cap A_\beta = \Phi,$$

Then called $\{ A_\alpha | \alpha\in I \}$ is a partition of dimension table A. Such as the teacher dimension table in the data warehouse,

© 2013. The authors - Published by Atlantis Press

each kind of teacher with different title is a partition of the teachers set.

Dimension class is a partition of dimension table A. Its significance is, according to certain criteria, to classify dimension table tuples.

B . The dimensional hierarchy

Let D is a dimension table, $A=\{a_1,a_2,\ldots\ldots,a_n,n\in I \text{ and } n\geq 1\}$ is a set of certain properties of D, the corresponding range of values is $Va_1,Va_2,\ldots\ldots,Va_n$, F is a fact table associated D, corresponding to the $Va_1,Va_2,\ldots\ldots,Va_n$ in F is the metrics set of $M_1,M_2,\ldots\ldots,M_n$（$n\in I$ and $n\geq 1$）,≤ is a binary relationship of set of $M=\{M_1,M_2,\ldots\ldots,M_n\}$ in a binary relationship, if ≤ meets:

- Reflexivity, any $M_i\in M(i\in[1,n]),M_i\leq M_i$ established,
- Anti-symmetry, any $M_i\in M(i,j\in[1,n])$, $M_i\leq M_j$ and $M_j\leq M_i=>M_i=M_i$ established,
- Transitivity, any $M_i,M_j,M_k\in M(i,j,k\in[1,n]),M_i\leq M_j$ and $M_j\leq M_k=>M_i\leq M_k$ established,

Claiming that ≤ is a partial order on the set M and the set A is dimension hierarchy which rely on the ≤.

The practical significance of the dimension hierarchy is, according to different extent, to gather metrics of the fact table which the dimension tables associated.

C . The apparent dimension hierarchy

According to the definition described in B of part II, the set $A=\{a_1,a_2,\ldots\ldots,a_n, n\in I \text{ and } n\geq 1\}$ is the dimension hierarchy, a collection of $M=\{M_1,M_2,\ldots\ldots,M_n\}$ is the metrics set, $m_{ik}\in M_i$, where $i\in[1,n],k\in I$ and $k\geq 1$. If it satisfies the conditions:

$$m_{jk}=\sum_{m_{ik}\in M_i} m_{ik}\text{, where } j=i+1,$$

Then set A is called the apparent dimension hierarchy.

The practical significance of the apparent dimension varies to determine the different extent, the value of the father-level metric data in the fact table is the integrated value of its sub-level. It should be noted that all sigma ∑ in this article does not refer to seek co-operation, but on behalf of all aggregation operations.

D . Hidden dimension hierarchy

Suppose the set $A=\{a_1,a_2,\ldots\ldots,a_n ,n\in I \text{ and } n\geq 1\}$ with properties of the dimension table D, there is $a_i\in A$, the corresponding range of values is Va_i,Va_i in accordance with an equivalent relation R can be divided into a set of equivalence class, $Va_i=\{[Va_{i1}]_R,[Va_{i2}]_R,\ldots\ldots,[Va_{im}]_R,m\in I \text{ and } m\geq 1\}$, a collection $M=\{M_1,M_2,\ldots\ldots,M_m\}$ is the metric collection Va_i corresponding to, $m_{ik}\in M_i,i\in[1,n],k\in I$ and $k\geq 1$. Suppose $Va_i' = \{[Va_{i1}]_R, [Va_{i1}]_R\cup[Va_{i2}]_R, \ldots\ldots, [Va_{i1}]_R\cup[Va_{i2}]_R\cup\ldots\ldots\cup[Va_{im}]_R, m\in I \text{ and } m\geq 1\}$, set $M'=\{M_1, M_1\cup M_2, \ldots\ldots, M_1\cup M_2\cup\ldots\ldots\cup M_m\}=\{MM_1, MM_2, \ldots\ldots, MM_m\}$, $mm_{ik}\in MM_i$, where $i\in[1,n]$, $k\in I$ and $k\geq 1$. If Va_i' defined in accordance with B of part II, and the condition is satisfied:

$$mm_{jk}=\sum_{m_{ik}\in M_i} m_{ik}+\sum_{m_{jk}\in M_j} m_{jk}\text{ ,where } j=i+1,$$

Then set Va_i' is hidden dimension hierarchy.

The hidden dimension hierarchy practical significance is determined the different degree of integration of the fact table metrics. Different from apparent dimension hierarchy, the father's level value is integrated from its sub-level.

III. The conditions of generating the apparent dimension hierarchy from the hidden dimension hierarchy

A . The certainty of property value of tuples in dimension table

Suppose property of dimensional table D is the set $A=\{a_1,a_2,\ldots\ldots a_n, n\in I \text{ and } n\geq 1\}$, then for any $a_i\in A$, there must be a definite Va_i corresponding to it, the set $Va_i=\{va_{i1},va_{i2},\ldots\ldots va_{ik}\}$ ($k\in I$ and $k\geq 1$) is the value range of dimension table a_i.

The condition indicates that property of dimensional table must have a corresponding value domain, by certain mapping, any tuple in the dimension table can determine the only property value.

B . Attributes in the dimension table has dimensional nature of the class

The attribute a_i of dimension table D corresponds to the range of values Va_i,if a binary equivalence relation R exist on Va_i, dimension table tuples set can be divided into equivalence classes, then the attributes in the dimension table has dimension class properties.

The condition means that the property value range in the dimension table can be divided into a finite number of subsets, in which tuples have common property feature.

C . The metric aggregation operation result corresponding to the deformed equivalence class of property value range of the dimensional table has a partial order

Set $s_1,s_2,\ldots\ldots s_n$（$n\in I$ and $n\geq 1$）. It is equivalence class corresponding to a particular property value range. To get the deformed equivalence class $s_1,s_1\cup s_2,\ldots\ldots,s_1\cup s_2\cup\ldots\ldots\cup s_n$, if they correspond to the metric set of $MM_2,\ldots\ldots MM_n$（$n\in I$ and $n\geq 1$）, metric $mm_{ik}\in MM_i$（$i\in[1,n],k\in I$ and $k\geq 1$）, f is a mapping function on the metric. The metric aggregation set corresponding to f is:

$$\{T|T=\sum_{m_{ik}\in M_i} f(m_{ik})+\sum_{m_{jk}\in M_j} f(m_{jk})\text{ , where } j=i+1\}$$

If ≤ is a partial order on T, the deformed equivalence class corresponding to partial order set T s_1, $s_1\cup s_2,\ldots\ldots$, $s_1\cup s_2\cup\ldots\ldots\cup s_n$ is the dimension hierarchy.

The conditions explain that, according to the basis of gathering requirements, mapping and aggregation operations on fact metric from equivalence classes of property value range of the dimensional table, if each obtained aggregation result and another is in partial order, Then equivalence class of property value range of the dimensional table can be identified as the dimension hierarchy.

IV. The implementation of generating the dominant dimension from the hidden dimensional hierarchy

A . Steps

(1) With the conditions described in A of part Ⅲ, ensure the values correctness of tuples property which wants to generate the dimension hierarchy.

(2) Based on the condition of B of part III and the definition of A of part II, determine the dimension class feature of property which wants to generate the dimension property hierarchy, it means dividing tuples in dimension tables on certain condition.

(3) Based on the condition of C of part III and the definition of B of part II, to determine the every equivalence class of value range of property which wants to generate the dimension property hierarchy and whether the result of the aggregation of the metric on every equivalence class is a partial order after some mapping process.

(4) With the definitions of C of part II and D of part II, to determine the feature of hidden dimension hierarchy of the generated dimension hierarchy from step (3).

(5) According to the functionality and performance of the system OLAP, to determine whether to form the hidden dimension hierarchy generated from step (4) to apparent dimension hierarchy. If the forecast can be identified that OLAP probability related to the apparent dimension hierarchy is very low, can choose to give up generation, and terminate the design of the dimension hierarchy. Otherwise, turn to steps (6) and (7).

(6) To define the Meta data in the system and describe new dimension hierarchy in the dimension table.

(7) To form the aggregation tables by the dimension hierarchy generating in step (5) for further use in OLAP

B . An Application

The CDCDH method is used in the establishment of data warehouse and decision analysis system design process in university. In the data warehouse, the teacher dimension table TeacherD holds Teacher ID, Department ID, Sex, Title ID, Education ID, Age ID properties, and design TeacherD with three sub-dimension elements Sub-TitleD, Sub-EducationD and Sub-AgeD. The table Sub-tileD generally divides into five categories Responsibility Professor, Professor, Associate Professor, Lecturer and Teaching Assistant respectively. The table Sub-EducationD can be divided into four categories Post Doctorate, Doctor, Master and Bachelor respectively. The table Sub-AgeD divides into Senior Teacher (50 years old), Middle-Aged Teacher (40 to 50 years old) and young teachers (40 years old). Obviously, all these are dimension hierarchies generating from mapping operations of the metric of value range. So can determine the hidden dimension hierarchies and generate apparent dimension hierarchies.

(1) Any tuple in the teacher dimension table has clear property value of title, educational background, age, etc.

(2) According to the same title, same age, same degree, tuples can be classified into each hierarchy in title sub-dimension table Sub-TitleD, degree sub-dimension table Sub-EducationD and age sub-dimension table Sub-AgeD.

(3) Define the mapping:

f: $x \mapsto$ Count(x), where $x \in$ {{Responsibility Professor}, {Professor}, {Associate Professor}, {Lecturer}, {Teaching Assistant}}

g: $x \mapsto$ Count(x), where $x \in$ {{Post Doctorate}, {Doctor}, {Master}, {Bachelor}}

h: $x \mapsto$ Count(x), where $x \in$ {{Senior Teacher}, {Middle-Aged Teacher}, {Young Teacher}}

Obviously.

A partial order is formed with the number of assistant teachers above> the number of lectures above >the number of vice professors above >the number of professors above >the number of responsibility professors above, which holds the hidden dimension hierarchy feature.

A partial order is formed with the number of bachelor degree above> the number of master degree > the number of doctorate above, which holds the hidden dimension hierarchy feature.

A partial order is formed with the number of Young teachers>the number of middle-aged teachers>senior teachers, which holds the hidden dimension hierarchy feature.

So the above hidden dimension hierarchy can be converted to apparent dimension hierarchy on demand for OLAP.

V . Conclusions

CDCDH method is introduced and a comprehensive discussion is given on dimension analysis and designs, which can avoid the performance decline caused by defects in design process and reduce the development cost. Of course CDCDH method has its shortcomings. In example B of part IV, the dimension table structure includes not only the dimension class attribute fields but also sub dimension hierarchy tables generating from dimension class, the larger storage space is needed. With the development of Mapreduce, a lot of storage space is shared on the network, OLAP efficient response will become an important research direction.

Reference

[1] Wang Shan, Data Warehousing and On-Line Analytical Processing. Beijing: Publishing House of Science,1999,pp.50-128.

[2] Leng Fangling, Bao Yubin, Yu Ge, Gao Wei, "Closed Data Cube Based on MapReduce",Journal of Computer Research and Development,vol.48, no.12, March 2011,pp .232-238.

[3] SHI Jingang, BAO Yubin, LENG Fangling, YU Ge, "Efficient Parallel Dwarf Data Cube Using MapReduce", Journal of Frontiers of Computer Science & Technology, vol.5,no.5,May 2011,pp. 398-409

[4] Xi Jian-qing, You Jin-guo, Tang De-you, Xiao Wei-ji, "A Parallel Closed-Cubing Algorithm Based on MapReduce", Journal of South China University of Technology(Natural Science Edition), vol.37,no.1, December 2009,pp.91-94.

[5] Leng Fangling, Bao Yubin, Gao Wei, Yu Ge, "MapReduce-based data aggregation algorithms Sciencepaper Online",vol.6,no.7,June 2011, pp.469-481.

.