The Description of Metadata of the Multidimensional Information Systems using Test Data

Maxim B. Fomin

Peoples' Friendship University of Russia (RUDN University)
6 Miklukho-Maklaya St., Moscow, 117198, Russian Federation

Email: fomin_mb@rudn.university

This paper examines the possibility of application of test data in the metadata description in information systems constructed on the basis of a multidimensional approach. In case of the description of the characteristics of the observed phenomenon using a large number of aspects, the multidimensional data cube, which is the basis of the information system, is characterised by high sparsity. It complicates the organization of data storage. This paper proposes clustered method of describing the data, which makes it possible to express the semantics of the subject domain. It is necessary to select the groups of members for dimensions that are semantically associated with the groups of members of other dimensions. The relationship between groups of members of different dimensions allows to identify clusters in the data cube, i.e. sets of cells that have similar properties and can be described in the same way. Clusters are used as the main element of information system data model. The problem of metadata description in the information system leads to the problem of setting the parameters of such clusters. Test data can be used in the process of describing the structure of a multidimensional cube. Such structures are data models that express the individual properties of the observed phenomenon. Test data can also be used in the process of testing possible methods of data analysis in multidimensional cube. In the process of development of a multidimensional information system can be used different methods to generate test data to suit the structure of clusters of cells in a multidimensional cube. The first method is applied when setting the values of measures that are semantically not related. The facts in this case are described by the Cartesian product of groups of values of measures. The second method is applied if values of measures correspond to different aspects of the same characteristic. The third method is applied if there is a correspondence between members and values of measures in the facts.

The publication was financially supported by the Ministry of Education and Science of the Russian Federation (the Agreement No 02.A03.21.0008).

Key words and phrases: multidimensional data models, test data, sparse data cube, set of possible member combinations, cluster of member combinations.

In: K. E. Samouilov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the VII Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Moscow, Russia, 24-Apr-2017, published at http://ceur-ws.org

Описание метаданных многомерной информационной системы с использованием тестовых данных

М. Б. Фомин

Российский университет дружбы народов
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия

Email: fomin_mb@rudn.university

В работе рассматривается возможность использования тестовых данных при описании метаданных в информационных системах, построенных на базе многомерного подхода. В случае многоаспектного описания характеристик наблюдаемого явления многомерный куб данных, лежащий в основе информационной системы, характеризуется большой разреженностью. Эффективным методом описания данных в этом случае, позволяющим выразить семантику предметной области, является кластерный метод. Для измерений, которые являются размерностями многомерного куба, выявляются группы их значений, которые семантически связаны с группами значений других измерений. Построение связей между группами значений разных измерений позволяет выявить в кубе данных кластеры – наборы ячеек, которые обладают сходными свойствами и могут быть описаны единым образом. Описание метаданных информационной системы сводится к заданию параметров таких кластеров. В процессе описания структуры многомерного куба могут использоваться тестовые данные – модели данных, выражающих отдельные свойства наблюдаемого явления. Тестовые данные могут применяться при тестировании возможных способов анализа данных многомерного куба. В процессе разработки многомерной информационной системы могут использоваться различные способы формирования тестовых данных, учитывающие особенности структуры кластеров ячеек многомерного куба. Первый способ применяется при задании показателей, которые семантически не связаны. Факты в этом случае описываются декартовым произведением групп значений показателей. Второй способ применяется в случае если значения показателей в фактах выражают разные аспекты одной характеристики. Третий способ применяется в случае если в фактах имеется соответствие между значениями показателя и значениями измерений.

Публикация подготовлена при финансовой поддержке Минобрнауки России (соглашение № 02.A03.21.0008).

Ключевые слова: многомерная модель данных, тестовые данные, разреженный куб, сочетание значений измерений, кластер сочетаний значений измерений.

1. Введение

В случае применения многомерного подхода описание метаданных информационной системы может быть сформировано посредством описания сочетаний значений характеристик наблюдаемого явления, определяющих значения анализируемых показателей. Аналитическим пространством информационной системы при таком подходе является многомерный куб данных, размерностями которого выступают измерения, соответствующие различным аспектам наблюдаемого явления. В случае многоаспектного анализа большого объема разнородных данных многомерный куб характеризуется высокой разреженностью и неравномерностью заполнения [1]. Описание аналитического пространства, выражающее семантику наблюдаемого явления, может быть произведено с использованием кластерного метода [2]. Суть метода заключается в объединении семантически схожих сочетаний значений измерений в кластеры сочетаний. Аналитическое пространство может быть представлено как множество таких кластеров. Описание структуры многомерного куба данных – сложная задача. Её важным элементом является тестирование

результатов описания, в процессе которого могут применяться тестовые данные – модели данных, имеющие небольшие размеры, представленные в соответствии со структурой метаданных информационной системы и выражающие отдельные свойства наблюдаемого явления.

2. Постановка задачи

При описании аналитического пространства многомерной информационной системы должны быть заданы параметры показателей и измерений.

Каждому аспекту анализа соответствует одно из измерений многомерного куба H. Полный набор измерений образует множество $D(H) = \{D^1, D^2, ..., D^n\}$, где D^i — i-е измерение, $n = \dim(H)$ — размерность многомерного куба. Каждое измерение задается множеством значений измерения $D^i = \left\{d_1^i, d_2^i, ..., d_{k_i}^i\right\}$, где i — номер измерения, k_i — количество значений измерения. В измерении могут быть заданы иерархии. Для задания иерархии все множество значений измерения должно быть разбито на уровни и между уровнями установлены иерархические связи в соответствии с правилами агрегации, действующими в отношении значений измерений. Для значений измерений могут быть заданы атрибуты, для каждого уровня в иерархии значений — свой набор атрибутов.

Многомерный куб данных может быть представлен как структурированный набор ячеек. Каждой ячейке c соответствует сочетание $c = \left(d_{i_1}^1, d_{i_2}^2, ..., d_{i_n}^n\right)$ значений измерений, по одному значению для каждого из измерений. В случае использования при анализе наблюдаемого явления большого набора разнотипных аспектов, не все возможные сочетания значений измерений задают значимые ячейки, то есть ячейки, соответствующие некоторому описываемому факту. Структуру аналитического пространства многомерной информационной системы определяет множество допустимых сочетаний значений измерений, соответствующее множеству значимых ячеек многомерного куба.

В некоторых сочетаниях значений измерений может возникать ситуация, когда одно или несколько измерений становятся семантически неопределенным в сочетании со значениями остальных измерений. При описании значимой ячейки многомерного куба, соответствующей описанной ситуации, для задания значения такого неопределенного измерения используется специальное значение «Не используется». Для значения «Не используется» не могут быть заданы значения атрибутов.

Количественно наблюдаемое явление характеризуется значениями показателей, заданными в значимых ячейках многомерного куба. Полный набор показателей образует множество $V(H) = \{v_1, v_2, ..., v_m\}$, где v_j — j-й показатель, m — число показателей в кубе. Между значениями измерений, задающими значимую ячейку, и показателями может возникать семантическое несоответствие. В этом случае в значимой ячейке могут быть заданы не все показатели из $V(H)$. Для задания в ячейке семантически не определенного показателя можно использовать специальное значение «Не используется».

Ячейки, обладающие сходными семантическими характеристиками, могут быть объединены в кластер. Кластеру ячеек соответствует кластер сочетаний значений измерений — множество сочетаний значений измерений, в котором для любой пары измерений значения этих измерений, имеющиеся в сочетаниях кластера, могут находиться в сочетании «каждый с каждым» со схожей семантикой. Набор значений измерения образует группу значений измерения в кластере, полный набор сочетаний кластера может быть получен при помощи операции декартова произведения,

в которой операндами являются группы значений измерений или специальное значение «Не используется», по одному операнду для каждого из измерений, заданных в многомерном кубе.

Объединение всех кластеров допустимых сочетаний значений измерений, которые могут быть сформированы, дает полное описание аналитического пространства многомерной информационной системы.

3. Формирование тестовых данных

Полноту описания метаданных многомерной информационной системы кластерным методом можно протестировать путем наполнения системы тестовыми данными, моделирующими свойства различных смысловых компонент наблюдаемого явления. В процессе такого моделирования должны быть заданы значения показателей для сочетаний значений измерений тех кластеров, которые соответствуют рассматриваемой смысловой компоненте. Анализ этих данных в разных разрезах позволяет сделать вывод о полноте описания метаданных системы.

Второй способ применения тестовых данных в процессе разработки многомерной информационной системы – отработка методики анализа хранящейся в системе информации. Процесс анализа состоит в преобразовании данных многомерного куба, в результате которого формируется небольшой объем данных, представляющий в удобном для анализа виде исследуемую характеристику. Преобразование должно быть описано на уровне метаданных как цепочка операций. На вход преобразования подается многомерный куб данных, который в процессе выполнения операций редуцируется в куб данных меньшей размерности и небольшого объема. На этапе отработки такой методики можно вместо полного куба данных использовать тестовые данные, в которых заложены те или иные свойства наблюдаемого явления.

В наблюдаемом явлении набору характеристик, которые описываются сочетанием значений измерений, может соответствовать несколько фактов. В информационной системе это свойство выражается в возможности задания нескольких наборов значений показателей для одного допустимого сочетания значений измерений.

Значения показателей в силу их семантических особенностей могут по-разному сочетаться между собой и с допустимыми сочетаниями значений измерений, входящих в кластер:

1) по правилу «каждый с каждым»;
2) в строгой привязке определенных значений одного показателя к определенным значениям другого показателя;
3) в строгой привязке определенных значений показателей к определенным значениям измерений допустимого сочетания.

Эти свойства фактов должны найти отражение в тестовых данных.

В первом случае показатели должны описывать семантически не связанные характеристики фактов. В качестве примера можно привести показатели «Место выдачи кредита» и «Период начисления процентов» для наблюдаемого явления «Кредитование». Для значений указанных показателей строгое соответствие между собой и с сочетаниями значений измерений кластера не устанавливается: значения показателей могут присутствовать в факте в любом сочетании. Соответствующие кластеру факты могут быть получены декартовым произведением трех операндов: допустимых сочетаний значений измерений кластера и значений двух показателей. Связи между значениями показателей «Место выдачи кредита» и «Период начисления процентов», формирующие факты кластера, приведены на диаграмме ниже (рис. 1).

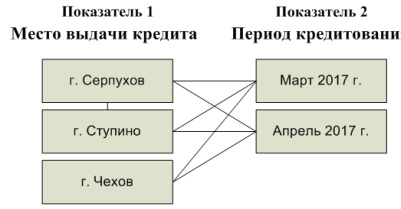

Рис. 1. Диаграмма связей между значениями показателей в кластере для случая описания несвязанных характеристик

Задание в факте значений одного показателя в строгой привязке к определенным значениям другого показателя появляется в случае если эти показатели выражают разные аспекты одной сущности. Такая ситуация может быть проиллюстрирована на примере показателей «Заемщик», «ИНН заемщика» и «Вид деятельности заемщика». Значения этих показателей должны присутствовать в факте в строгом соответствии друг с другом, построение моделей фактов с использованием декартова произведения привело бы к бессмысленным фактам. Связи между значениями показателей со строгим взаимно однозначным соответствием представлены на диаграмме ниже (рис. 2).

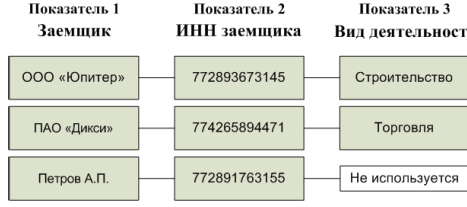

Рис. 2. Диаграмма связей между значениями показателей в кластере для случая описания характеристик со строгим соответствием значений

В одном из фактов на диаграмме показатель «Вид деятельности» принимает значение «Не используется». Такой эффект возникает вследствие того, что этот показатель семантически не определен в случае если значение показателя «Заемщик» в факте соответствует физическому лицу.

Случай привязки значений показателей к значениям измерений в допустимых сочетаниях может быть проиллюстрирован примером, в котором факты кластера задаются характеристиками «Заемщик», «ИНН заемщика» и «Место жительства заемщика». Здесь, в отличие от ситуации, описанной в предыдущем примере, характеристика факта «Заемщик» выведена из состава показателей и представлена в виде измерения аналитического пространства «Клиент банка». В этом случае значения показателей «ИНН заемщика» и «Место жительства заемщика» в факте соответствуют значениям атрибутов «ИНН» и «Место жительства», заданных для

значений измерения «Клиент банка». Связи между значениями показателей в случае соответствия со значениями измерений в факте представлены на диаграмме ниже (рис. 3).

Рис. 3. Диаграмма связей между значениями показателей в кластере для случая установления соответствия со значениями измерений

При моделировании тестовых данных многомерных информационных систем необходимо использовать способы задания значений показателей в кластере моделей фактов, соответствующие описанным особенностям фактов и отражающим эти особенности свойствам сочетаемости значений показателей. Можно предложить три способа задания значений показателей в кластере фактов.

Первый способ применяется в случае задания значений показателей, которые семантически не связаны в фактах, относящихся к описываемому фрагменту наблюдаемого явления. Каждый из таких показателей задается группой значений показателя. При формировании фактов используется декартово произведение, в котором операндами являются значения из групп значений разных показателей и допустимые сочетаниями значений измерений кластера.

Второй способ применяется в случае если в моделируемых фактах значения нескольких показателей выражают разные аспекты одной характеристики наблюдаемого явления. Такие показатели задается группами значений показателей с установлением порядка следования значений в этих группах. Должно выполняться дополнительное требование: для всех показателей количество элементов в группах – одинаковое. При формировании фактов используется поэлементное соединение значений из групп, относящихся к разным показателям, в наборы значений в соответствии с порядком следования значений в группах. После этого выполняется декартово произведение сформированных наборов со значениями оставшихся показателей и с допустимыми сочетаниями значений измерений кластера.

Третий способ применяется в случае если в моделируемых фактах имеется соответствие между значениями показателя и значениями одного или нескольких измерений. В этом случае показателю присваивается значение, однозначно соответствующее значению одного из измерений или значениям нескольких измерений в допустимом сочетании значений измерений, заданном в факте. Механизм установления соответствия между значениями показателей и измерений может быть реализован через описание атрибутов для значений связанных измерений: для измерения должен быть задан атрибут, а для значений измерений заданы значения этого атрибута. Значения такого атрибута должны использоваться в качестве значений связанного показателя. В случае если в допустимом сочетании значений измерений, соответствующем факту, рассматриваемое измерение имеет значение «Не используется» или для значения этого измерения не задан рассматриваемый атрибут, соответствующий показатель приобретает значение «Не используется».

4. Выводы

В работе был рассмотрен способ задания тестовых данных, которые моделируют различные смысловые компоненты наблюдаемого явления. Тестовые данные могут быть использованы при анализе семантической обоснованности и полноты описания метаданных многомерной информационной системы и при отработке методики анализа данных, представленных в виде многомерного куба. В процессе задания тестовых данных должны учитываться особенности моделируемой смысловой компоненты наблюдаемого явления. Эти особенности должны выражаться в выборе способа задания сочетаний значений характеристик наблюдаемого явления, включенных в многомерный куб данных.

Литература

1. E. Thomsen, OLAP Solution: Building Multidimensional Information System. Willey Computer Publishing, 2002. ISBN 0-471-40030-0.
2. M. B. Fomin, Cluster Method of Description of Information System Data Model Based on Multidimensional Approach, Distributed Computer and Communication Networks. Springer, 2016. ISBN 978-3-319-51917-3, pp. 657–668.