DOI: 10.32347/2412-9933.2020.42.93-101
УДК 004:728+827

Гончаренко Тетяна Андріївна
Кандидат технічних наук, доцент, доцент кафедри інформаційних технологій, *orcid.org/0000-0003-2577-6916*
Київський національний університет будівництва і архітектури, Київ

КЛАСТЕРНИЙ МЕТОД ФОРМУВАННЯ МЕТАДАНИХ БАГАТОВИМІРНИХ ІНФОРМАЦІЙНИХ СИСТЕМ ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ ГЕНЕРАЛЬНОГО ПЛАНУВАННЯ

Анотація. *Розроблено спосіб формування метаданих багатовимірної інформаційної системи шляхом сполучення класифікаційних схем. Кожна класифікаційна схема являє собою ієрархію значень вимірів, що належить до окремої структурної компоненти генерального плану (ГП). В основі методу лежить виявлення груп значень вимірів, які пов'язані з групами значень інших вимірів. Групи значень різних вимірів використовуються для побудови кластерів поєднань значень вимірів. Сполучення кластера формуються декартовим добутком груп значень вимірів. Метадані інформаційної системи представлені у вигляді множини допустимих поєднань значень вимірів, які формуються як набір кластерів. Для вирішення цього складного завдання ГП розглядається як набір структурних компонентів. З повного набору вимірювань інформаційної системи виокремлюються набори вимірювань, що семантично пов'язані зі структурними компонентами ГП. Семантичні зв'язки, виявлені в процесі аналізу структурної компоненти, дають змогу побудувати ієрархію груп значень вимірів і представити їх сукупність у вигляді графа – класифікаційної схеми, пов'язаної зі структурною компонентою. В інформаційних системах з багатоаспектним описом предметної області куби даних характеризуються великою розрідженістю, що ускладнює формування метаданих. Класифікаційні схеми описують окремі аспекти метаданих, пов'язані з окремими структурними компонентами ГП. Сполучення класифікаційних схем дає можливість отримати повний опис метаданих. Використання класифікаційних схем допомагає розбити задачу опису структури аналітичного простору багатовимірної інформаційної системи на більш прості задачі аналізу його окремих структурних компонент. Сполучення класифікаційних схем, що належать до різних структурних компонентів, дає можливість сформувати метадані інформаційної системи. У метаданих центральне місце посідає множина допустимих поєднань значень вимірів.*

Ключові слова: багатовимірна інформаційна система; генеральне планування; багатовимірний куб даних; розріджений куб даних; класифікаційна схема

Актуальність та аналіз проблеми

В інформаційній системі, в якій показники, що характеризують ГП, представлені в багатовимірній формі, розмінностями куба даних є вимірювання. Кожен вимір відповідає деякому аспекту аналізу спостережуваного явища. У разі якщо система містить великий обсяг семантично різнорідних даних, багатомірний куб даних характеризується високою розрідженістю і нерівномірністю заповнення [1]. Модель даних інформаційної системи формується згідно з таким принципом: кожна значуща комірка багатовимірного куба відповідає деякому факту. Для ефективного опису структури багатовимірного куба можна використати кластерний метод. Цей метод базується на семантичному аналізі сполучуваності значень різних вимірів у значущих комірках куба [2]. Він допомагає описати метадані інформаційної системи у вигляді множини допустимих поєднань значень вимірів. Допустимі поєднання ставляться у відповідність значущим коміркам багатовимірного куба.

Постановка задачі

У разі якщо багатовимірна інформаційна система створюється для опису семантично різнорідних фактів, і структура аналітичного простору містить велику кількість вимірювань, багатовимірний куб даних характеризується значною розрідженістю, яка має бути відображена в моделі даних [3 – 9]. У цій ситуації при описі множини допустимих поєднань виникає складна задача аналізу сполучуваності значень всіх вимірювань куба в сукупності. Ця задача може бути спрощена, якщо процес ГП допускає поділ на набір структурних компонент, кожна з яких має свої аспекти аналізу. Облік семантики має важливе значення при побудові моделі даних [10]. Поділ на набір структурних

компонент допоможе виокремити в аналітичному просторі набори вимірювань, асоційовані зі структурними компонентами, і розглядати сполучуваність значень вимірів в кожному наборі вимірювань окремо.

Мета статті

Метою дослідження є розроблення методу побудови множини допустимих поєднань значень вимірів багатовимірного куба, що складається з таких етапів:

1) декомпозиція об'єкта, який описує інформаційна система, на структурні складові;

2) аналіз сполучуваності значень вимірів, що характеризують ці структурні складові;

3) побудова класифікаційних схем, що містять опис допустимих поєднань значень вимірів окремо для кожної структурної складової;

4) з'єднання сполучень, взятих з різних класифікаційних схем, в множині допустимих поєднань значень вимірів багатовимірного куба в сукупності.

У процесі виконання описаного вище алгоритму характеристики об'єкта і зв'язки між ними треба розглядати з позиції класифікації, яка відображала б семантику об'єкта. Як характеристики виступають вимірювання куба даних. Класифікацію характеристик можна виконати з використанням ієрархічного принципу. В цьому випадку виявлення властивості можуть бути представлені у формі зв'язного ациклічного графа. Характеристики об'єкта поділяються за ознакою значущості і розподіляються за різними рівнями ієрархії графа. Після формування ієрархії характеристик можна переходити до побудови графа, використовуючи при цьому попарний аналіз сполучуваності значень вимірів, що відповідають характеристикам, розташованими в ієрархії одна під одною.

Виклад основного матеріалу

Опис розрідженого куба даних з використанням поєднань значень вимірів

Кожному аспекту аналізу об'єкта, для опису якого розробляється багатовимірна інформаційна система, відповідає один з вимірів багатовимірного куба H. Повний набір вимірювань утворює множину

$$D(H) = \{D^1, D^2, .., D^i, .., D^n\},$$

де D^i – *i-й* вимір, $n = \dim(H)$ – розмірність багатовимірного куба.

Вимірювання задається множиною значень вимірювання:

$$D^i = \{d_1^i d_2^i, .., d_{k_i}^i\},$$

де k^i – число значень i-го вимірювання. Значення вимірювання D^i вибираються з множини позицій класифікатора, який відповідає тому аспекту спостережуваного явища, який пов'язаний з вимірюванням D^i.

Багатовимірний куб даних є структурованим набором комірок. Кожній комірці c багатовимірного куба може бути зіставлене сполучення значень вимірів:

$$c = (d_{i_1}^1 d_{i_2}^2, .., d_{i_n}^n),$$

по одному значенню для кожного з вимірів [11]. У разі розрідженого куба не всі можливі сполучення значень вимірів відповідають значущим коміркам куба, тобто описують існуючі факти.

Якщо багатовимірний куб містить семантично різнорідні дані, можлива ситуація, коли значення деяких вимірювань не можуть бути задані в поєднанні з наявним набором значень інших вимірів. У такій ситуації при описі значущого параметра багатовимірного куба значення деяких вимірювань не можуть бути визначені. Для визначення значень цих семантично невизначених вимірювань доцільно застосовувати спеціальне значення «Не використовується» [11]. Структуру багатовимірного куба даних інформаційної системи в цьому випадку можна описати як множину допустимих поєднань значень вимірювань. У сполученнях цієї множини можуть використовуватися значення, взяті з класифікаторів, відповідних вимірювань, а також спеціальне значення «Не використовується». Для позначення множини допустимих сполучень значень вимірів будемо використовувати абревіатуру «МДСЗ».

Процес ГП характеризується значеннями показників, заданими в значущих комірках багатовимірного куба. Повний набір показників утворює множину

$$V(H) = \{v_1, v_2, .., v_j, .., v_p\},$$

де v_j – j-й показник; p – число показників в гіперкубі.

Значущою коміркою можуть бути задані не всі показники з $V(H)$. Така ситуація виникає у випадку семантичної невідповідності між значеннями вимірювань, які задаються коміркою, і деякими показниками.

Для опису МДСЗ для кожної значущої комірки c потрібно задати свою множину $V(c) = \{v_1, v_2, .., v_{p_c}\}$, що складається з певних в цій комірці показників, $1 \le p_c \le p$. Для опису в комірку c показників, що не входять в множину $V(c)$, будемо застосовувати спеціальне значення «Не використовується». Повинно виконуватися правило: множина показників $V(c)$, заданих в значущій комірці c, не може бути порожньою. Опис показників в значущих комірках багатовимірного куба, які відповідають сполученням значень вимірів, що не входять в МДСЗ, не має сенсу.

Використання кластерного методу опису структури багатовимірного куба даних

Структура МДСЗ описує семантику процесу ГП, інформація про який міститься в багатовимірному кубі даних. Виявити структуру багатовимірного куба даних може допомогти класифікаційний підхід [12 – 15]. Для встановлення зв'язків між елементами різних розмірностей можуть бути використані непараметричні методи статистичного аналізу [16]. Стислий опис МДСЗ, що враховує семантику, може бути отримано за допомогою кластерного методу, який засновано на аналізі попарних зв'язків між значеннями вимірювань [17]. Кластерний метод допомагає виявити групи значень вимірів. Група $G_j^i = \{d_1^i, d_2^i, .., d_{m_j}^i\}$ значень i-го вимірювання включає m_j значень $(1 \le m_j \le k_j)$, де j – номер групи, що містить значення вимірювання, які «однаково» поєднуються в МДСЗ зі значеннями з деяких груп інших вимірів.

За допомогою семантичного аналізу процесу ГП можна виявити пов'язані групи значень в різних вимірах. Кластер сполучень K – це множина поєднань значень вимірів, які можуть бути отримані за допомогою операції декартового добутку, в якому операндами є групи значень вимірів або спеціальне значення «Не використовується», по одному операнду для кожного з вимірювань, що використовуються в кластері: $G_1, G_{2\times..\times}G_n$. Кластери поєднань можуть бути використані при описі МДСЗ.

В процесі ГП можна виокремити смислові компоненти, які відрізняються одна від одної. В цьому випадку можна сформувати підмножини поєднань, кожне з яких відповідає своїй смисловій компоненті. Підмножина сполучень є об'єднанням кластерів сполучень. Воно може бути побудовано як результат аналізу сполучуваності характеристик спостережуваного явища, відповідних деякій його смисловій компоненті. Технічно характеристики відображаються в кластерах у вигляді значень вимірів багатовимірного куба.

Кластерний метод дає змогу отримати опис МДСЗ для багатовимірного куба H шляхом виконання таких кроків.

1. У структурі ГП мають бути виокремлені N семантичних компонент. Цим компонентам повинні бути співставлені підмножини поєднань:

$$Q_k, k = 1, .., N.$$

Має бути побудовано вираз для множини допустимих поєднань МДСЗ (H), в якому підмножини Q_k зв'язані за допомогою теоретико-множинних операцій об'єднання, перетину і доповнення.

2. У кожній підмножині Q_k мають бути виокремлені шари вимірювань:

$$L^i = \{D^{j_1}, D^{j_2}, .., D^{j_l}\}$$

де $i = 1, .., m_k$ – номер шару в підмножині; m_k – кількість шарів; j_i – номер вимірювання в шарі; $j = 1, .., l$; l – кількість вимірювань в i-му шарі.

Шар вимірювань – це набір вимірювань, сполучуваність значень яких здебільшого не залежить від того, які значення в поєднаннях приймають вимірювання, що не входять в шар. Якщо відомі підмножини сполучень значень вимірів для кожного з шарів вимірювань, то підмножина поєднань Q_k може бути отримана за допомогою декартового добутку:

$$Q = \text{МДСЗ}(L^1) \times \text{МДСЗ}(L^2) \times..\times \text{МДСЗ}(L^m).$$

де $\text{МДСЗ}(L^i)$ – множина допустимих сполучень значень вимірювань i-го шару.

3. Для кожного з шарів L^i підмножини Q^k підмножина допустимих поєднань $\text{МДСЗ}(L^i)$ має бути представлена у вигляді набору кластерів сполучень в шарі. Кожен кластер в шарі визначається набором груп значень вимірів G_j^k:

$$K = \{G_1^{j_1} G_2^{j_2}, .., G_l^{j_l}\},$$

де k – номер вимірювання в шарі; j – номер групи, $j = 1, .., l$.

Сполучення кластера можуть бути отримані декартовим добутком груп значень вимірів (або спеціального значення «Не використовується» замість групи), по одній групі для кожного з вимірів шару:

$$\text{МДСЗ}(K) = G_1^{j_1}, G_2^{j_2} \times..\times G_l^{j_l}.$$

На рис. 1 представлена діаграма, яка описує взаємозв'язок структурних елементів МДСЗ.

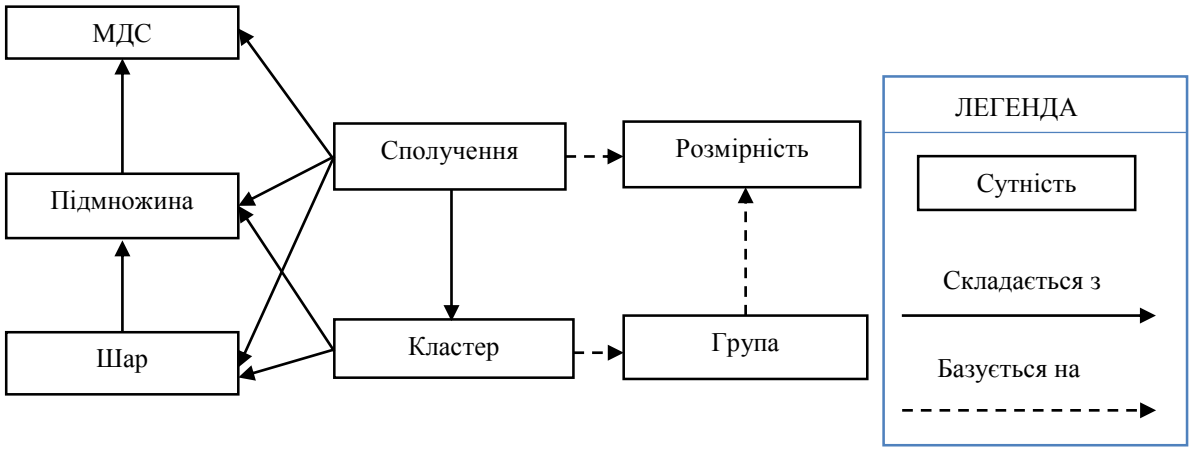

Рисунок 1 – Діаграма структури множини допустимих поєднань значень вимірів

Можна розглянути два типових випадки розбиття процесу ГП на смислові компоненти та подання МДСЗ (H) з використанням декількох підмножин. Перший – коли при аналізі різних смислових компонент виникають різні декомпозиції вимірювань на шари, другий – коли є простий спосіб побудови підмножини, що описує МДСЗ з надлишком, і ефективний спосіб опису сполучень, які мають бути виключені з цієї підмножини, щоб скоротити його до МДСЗ.

У першому випадку декомпозиції ГП на l смислових компонент відповідає об'єднання підмножин сполучень значень (ПСЗ) вимірів:

$$\text{ПСЗ}\ (H) = Q_1 \cup Q_2 \cup..\cup Q_l.$$

Внаслідок семантичних відмінностей цих смислових компонент, множини вимірів у різних підмножин можуть бути по-різному розбиті на шари:

$$D(H) = L_i^1 \cup L_i^2 \cup..\cup L_i^{m_i},$$

де $i = 1..l$ – номер розбиття; m_i – число шарів в i-му розбитті. Кожна підмножина Q_i формується відповідно зі своїм розбиттям множини вимірів на шари.

У другому випадку множина допустимих поєднань представляється у вигляді різниці підмножин:

$$\text{МДСЗ}(H) = R \backslash Q,$$

де R – множина сполучень, описана з надлишком (підмножина, яка скорочується); Q – множина виключних поєднань.

Підмножина, яка скорочується, може бути сформована з використанням такого правила: в неї включаються поєднання, отримані декартовим добутком всіх значень вимірів, доповнені набором сполучень, що містять значення «Не використовується» для деяких вимірювань, з виключенням тих поєднань, які можуть бути отримані заміною спеціального значення «Не використовується» на допустиме значення. Такий підхід може бути використаний у разі, якщо множина МДСЗ (H) має складну структуру і відомий простий алгоритм формування підмножини Q.

Опис спостережуваного явища набором класифікаційних схем

Процес опису властивостей ГП в рамках багатовимірної моделі даних з позицій семантики полягає у виявленні класифікаційних ознак (вимірювань багатовимірного куба) і встановленні зв'язків між ними. При цьому ГП не розглядається як багатокомпонентний об'єкт, а класифікаційні ознаки, що ранжуються, не розрізняють на головні і другорядні. Встановлення зв'язків між вимірами проводиться шляхом пошуку відповідності між їх значеннями. У разі великого числа вимірювань це складне завдання, недоліки такого підходу можуть бути усунені введенням в модель даних інформаційної системи додаткових об'єктів – класифікаційних схем характеристик (КСХ) ГП.

Визначимо для КСХ такі вимоги:

1. При задаванні КСХ має враховуватися компонентна структура ГП. Якщо ГП семантично може бути розділено на окремі структурні складові, для кожної з яких може бути обрано свій набір аспектів аналізу, для кожної такої складової має зіставлятися КСХ. Процедура побудови КСХ повинна базуватися на виявленні та аналізі відповідних обраним аспектам аналізу характеристик. Характеристиками мають бути зіставлені вимірювання багатовимірного куба.

2. КСХ ГП мають бути побудовані за ієрархічним принципом. Серед характеристик, що належать до КСХ, має бути встановлено ранжування, що виокремлює вимірювання, які більшою і меншою мірою висловлюють сенс структурної складової ГП, яка порівняна КСХ. Має бути обрано головний вимір, що найбільшою мірою відображає семантику відповідної КСХ структурної складової. З інших вимірів, включених в КСХ, які з семантичної точки зору підпорядковані головному виміру і висловлюють окремі властивості структурної складової ГП, повинна бути сформована ієрархія характеристик. Має бути реалізовано такий принцип: значення головного вимірювання виражають найбільш значущі властивості ГП; значення вимірювань, що лежать нижче за ієрархією по відношенню до головного, описують підлеглі властивості, уточнюючі сенс значень головного вимірювання.

3. При побудові ієрархій характеристик ГП КСХ повинна бути можливість опису значень головного вимірювання окремо або за групами значень, оскільки різні значення можуть бути пов'язані з різними аспектами семантики структурної складової ГП. Для значень головного вимірювання, що мають таке семантичне розходження, повинні бути побудовані різні ієрархії характеристик.

4. В ієрархії характеристик, яка є в КСХ, має бути наявна інформація про те, який набір показників кількісно описує ГП в разі вибору конкретних значень вимірів, наявних в ієрархії.

В процесі розроблення інформаційної системи класифікаційні схеми можуть взяти на себе роль джерела класифікаційної інформації про ГП. При цьому семантично КСХ пов'язана зі структурною складовою ГП і може бути джерелом інформації про характеристики структурної складової, представленої в ієрархічній формі. Технологічно КСХ пов'язана з вимірами багатовимірного куба даних, а отже, може бути шаблоном при побудові метаданих багатовимірної інформаційної системи.

Подання класифікаційної схеми у вигляді дерева поєднань

Класифікаційна схема характеристик ГП – об'єкт багатовимірної інформаційної системи, описує структурну складову ГП і містить такі дані:

– набір вимірювань, включених в класифікаційну схему;

– набір значень цих вимірів, включених в класифікаційну схему;

– головний вимір, вибраний в наборі вимірювань КСХ;

– набір показників, включених в класифікаційну схему;

– дерево поєднань значень вимірів КСХ, що задає ієрархію характеристик, включених в КСХ.

Ієрархічний принцип побудови КСХ реалізується в структурі дерева поєднань значень вимірів КСХ. Дерево поєднань КСХ може бути побудовано як результат семантичного аналізу структурної складової спостережуваного явища. Дерево можна визначити шляхом опису процедури його побудови. Побудова дерева має здійснюватися рухом від кореня дерева, в якому задані групи значень ключового вимірювання, вниз за рівнями ієрархії з додаванням в дерево на кожному кроці групи значень вимірів, що розкриває сенс значень вимірювання попереднього рівня ієрархії. При цьому на наступний рівень має бути додана група, що належить до вимірювання, найбільшою мірою пов'язаного зі значеннями вимірювання попереднього рівня. Як наслідок: в різних гілках дерева на шляхах від кореня до листя можуть виникати різні послідовності вимірювань КСХ.

При обході дерева КСХ від значень головного вимірювання вниз за рівнями ієрархії, значення вимірювань, розташовані на цих рівнях, у міру обходу висловлюють все менш значущі властивості ГП. Тим самим в дереві КСХ установлюється ранжування характеристик ГП.

В результаті виконання описаного вище алгоритму відбувається побудова дерева поєднань значень вимірів КСХ, що володіє структурою, для якої виконуються такі правила:

1. Коренем дерева є вузол «Ключовий вимір».

2. Дерево являє собою ієрархічну структуру, в якій рівні задаються чергуванням вузлів типу «Група значень вимірювання» і вузлів типу «Вимірювання». При цьому групи значень вимірів мають бути задані у вимірах, відповідних вузлів, розташованих в дереві на один рівень вище за ієрархією.

3. Листям дерева є вузли типу «Група значень вимірювання».

4. Вузлу типу «Група значень вимірювання» (крім вузла, що є листом дерева) має відповідати один вузол типу «Вимірювання» на розташованому нижче рівні ієрархії дерева. Вузлу типу «Вимірювання» може відповідати один вузол або кілька вузлів типу «Група значень вимірювання» на розташованому нижче рівні ієрархії дерева.

5. На шляху від кореня до листа кожен вимір може зустрічатися не більше одного разу.

Кожен шлях від кореня дерева сполучень до листа містить певний набір груп значень різних вимірів. Це означає, що шлях задає кластер сполучень значень вимірювань КСХ. Для формування повного набору кластерів сполучень потрібно обійти все дерево. У процесі обходу дерева в ширину число формованих кластерів збільшується кожного разу, коли на деякому рівні ієрархії зустрічається кілька груп значень, що належать до одного вузла типу «Вимірювання». Якщо на шляху від кореня дерева до листа відсутній деякий вимір, що є в структурі КСХ, то цей вимір має приймати значення «Не використовується» в кластері, який відповідає розглянутому шляху.

Кластери поєднань значень вимірів для КСХ, дерево поєднань значень вимірювань якої представлено на рис. 2, наведено у таблиці.

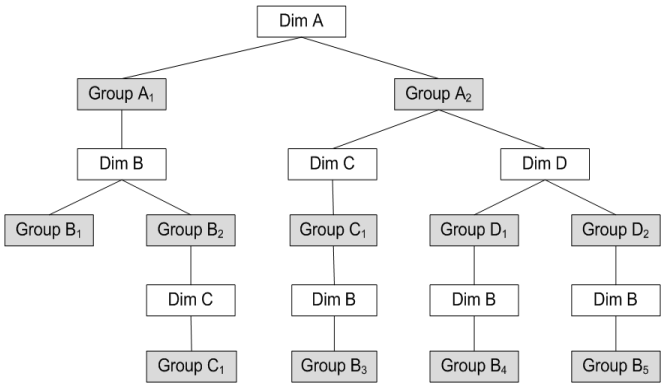

Рисунок 2 – Дерево поєднань значень вимірювань

Таблиця – ***Кластери поєднань значень вимірів класифікаційної схеми***

№	Вимір A	Вимір B	Вимір C	Вимір D
1	A1	B1	Не використовується	Не використовується
2	A1	B2	C1	Не використовується
3	A2	B3	C2	Не використовується
4	A2	B4	Не використовується	D1
5	A2	B5	Не використовується	D2

З позицій семантики кожен кластер, який є в дереві поєднань класифікаційної схеми, відповідає своєму набору властивостей спостережуваного явища. В інформаційній системі ці властивості описуються значеннями деякого набору показників, включених в класифікаційну схему. У різних кластерах можуть бути задані різні набори показників. Інформація про те, які показники задані в кластері поєднань КСХ, має бути описана в дереві поєднань значень вимірювань КСХ у вигляді атрибутів листя дерева поєднань.

Формування структури багатовимірного куба даних з використанням класифікаційних схем

Важливою властивістю КСХ є можливість використовувати сполучення значень вимірів, які в них описані, при формуванні метаданих інформаційної системи. Набір вимірювань багатовимірного куба формується виходячи з такого принципу: в ньому мають бути наявні всі характеристики, від яких можуть залежати показники, використовувані при аналізі ГП. Побудова МДСЗ на такому «широкому» наборі вимірювань багатовимірного куба є складним завданням. Вирішити це завдання допомагає використання КСХ, які відіграють класифікаційну функцію по відношенню до МДСЗ.

У кожній з КСХ, що належить до ГП, вже міститься інформація про сполучуваність значень тієї частини вимірів, які пов'язані з окремими властивостями ГП. Завдання полягає в тому, щоб правильно поєднати сполучення з різних КСХ разом у великій кількості допустимих поєднань. При з'єднанні сполучень двох КСХ може виникнути одна з таких ситуацій:

– вимірювання, включені до першої КСХ, відсутні в другій і навпаки;

– набори вимірювань, наявні в КСХ, частково перетинаються.

У першій ситуації сполучуваність значень вимірів однієї КСХ не залежить від значень вимірів іншої КСХ. Така ситуація відповідає випадку, коли МДСЗ, що описує структуру багатовимірного куба даних, розбита на шари. Для кожної з КСХ в цьому випадку МДСЗ (КСХ) є описом одного із шарів МДСЗ (H).

У ситуації перетину наборів вимірювань, що належать до двох КСХ, при формуванні сполучень МДСЗ (H) виникає задача продовження сполучень однією КСХ сполученнями іншої КСХ з частковим перекриттям за вимірюваннями. Ця задача може не мати однозначної відповіді. Вибір правильного варіанта побудови МДСЗ (H) в описаній ситуації має проводитися аналітиком на основі семантичного аналізу. При цьому мають бути вирішені такі питання:

– якщо значення вимірювань в поєднаннях з різних КСХ в області перетину вимірювань збігаються, то чи вважати такі поєднання продовженням один одного, або вважати, що кожне з них породжує своє поєднання в МДСЗ (H);

– якщо для деякого поєднання з однією КСХ є кілька продовжень в іншій КСХ, то який з можливих варіантів продовження вибрати при формуванні поєднання в МДСЗ (H).

Висновок

Метадані багатоаспектної багатовимірної інформаційної системи, спроєктованої з використанням кластерного методу, мають структуру розрідженого і нерівномірно заповненого багатовимірного куба. Формування таких метаданих є складним завданням, вирішити яке можна, розглядаючи ГП, який описує інформаційна система у вигляді сукупності структурних складових. Кожній структурній складовій відповідає класифікаційна схема, дані якої можуть бути представлені у вигляді множини допустимих сполучень значень вимірів, пов'язаних з характеристиками цієї структурної складової. Класифікаційні схеми, у порівнянні з метаданими інформаційної системи загалом, описують вузький набір властивостей ГП і представляють характеристики цих властивостей в ієрархічній формі. Вирішити завдання побудови класифікаційних схем допомагає семантичний аналіз характеристик структурних складових ГП кожної структурної складової окремо. Обмежений набір характеристик в КСХ спрощує процес формування ієрархії значень характеристик.

В результаті з'являється можливість виявлення внутрішньої структури багатовимірного куба даних. Підмножини вимірювань, включених в різні класифікаційні схеми, частково перетинаються. Стикування сполучень з різних КСХ відповідно до значень вимірів, що лежать в області перетину, дає змогу відновити структуру багатовимірного куба даних інформаційної системи. Ця процедура має бути виконана за участю аналітика, який приймає рішення про вибір способу продовження поєднання у випадку багатозначності.

У разі розробленя великої багатоаспектної багатовимірної інформаційної системи використання кластерного підходу для опису множини допустимих сполучень значень вимірів допомагає забезпечити компактність при визначенні метаданих і висловити семантику аналізованого ГП. В основі запропонованого підходу лежить виявлення зв'язків між вимірами, які відображають властивості ГП, і формування груп значень вимірів, елементи яких об'єднані схожою поведінкою по відношенню до цих зв'язків.

Список літератури

1. *Thomsen E. OLAP Solution: Building Multidimensional Information System. NY, Willey Computer Publishing, 2002, 688 p.*
2. *Viktor Mihaylenko, Tetyana Honcharenko, Khrystyna Chupryna, Yurii Andrashko, Svitlana Budnik, Modeling of Spatial Data on the Construction Site Based on Multidimensional Information Objects in 'International Journal of Engineering and Advanced Technology (IJEAT)', ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019, Page No. 3934-3940. URL: https://www.ijeat.org/wp-content/uploads/papers/v8i6/F9057088619.pdf*
3. *Hirata, C.M., Lima, J.C. Multidimensional cyclic graph approach: representing a data cube without common sub-graphs. Information Sciences, 2011, Vol. 181, P. 2626–2655, DOI: 10.1016/j.ins.2010.05.0*
4. *Salmam F.Z., Fakir M., Errattahi R. Prediction on OLAP data cubes. Journal of Information & Knowledge Management. 2016. Vol. 15. No. 2. P. 449–458. DOI:10.1142/S0219649216500222*
5. *Fu L.: Efficient evaluation of sparse data cubes. In: Li Q., Wang G., Feng L. Advances in Web-Age Information Management, vol. 3129 – WAIM 2004. Heidelberg, Springer, 2004. P. 336–345. DOI: 10.1007/978-3-540-27772-9_34*
6. *Romero O., Pedersen T.B., Berlanga R., Nebot V., Aramburu M.J., Simitsis A.: Using semantic web technologies for exploratory OLAP: A survey. IEEE Transactions on Knowledge and Data Engineering. 2015. Vol. 27. No.2. P. 571–588. DOI: 10.1109/TKDE.2014.2330822*
7. *Salmam F.Z., Fakir M., Errattahi R. Explanation in OLAP data cubes. Journal of Information Technology Research. 2014. Vol. 7. No. 4. P. 36–78. DOI: 10.4018/jitr.2014100105*
8. *Orlov Y., Gaidamaka Y., Zaripova E. Approach to estimation of performance measures for SIP server model with batch arrivals. In: Vishnevsky V., Kozyrev D. Distributed Computer and Communication Networks. DCCN 2015, vol 601. Cham, Springer, pp. 141–150. DOI: 10.1007/978-3-319-30843-2_15*
9. *Висков А.В., Фомин М.Б. Моделирование аналитических измерений в многомерных базах данных // Вестник Иркутского государственного технического университета. 2012. Т. 63. № 4. С. 15–19.*
10. *Gomez L.I., Gomez S.A., Vaisman A.A. generic data model and query language for spatiotemporal OLAP cube analysis. In: Rundensteiner, E., Markl, V., Manolescu, I., Amer-Yahia S., Naumann F., Ari I. Proceedings of the 15-th International Conference on Extending Database Technology – EDBT 2012. New York, ACM, 2012, P. 300–311.*
11. *Фомин М.Б. Описание метаданных многомерных информационных систем с использованием кластерного метода // Вестник Иркутского государственного технического университета, 2017, Т 21, № 7. С. 78–86. https:doi.org/10.21285/1814-3520-2017-7-78-86*
12. *Гончаренко Т.А. "Застосування BIM-технології для створення інформаційної моделі території під забудову", Управління розвитком складних систем, № 33, с. 138–145, 2018. [Видання включено до НБД: BASE; Index Copernicus].*
13. *Гончаренко Т.А. "Об'єктно-орієнтоване моделювання просторових об'єктів генерального планування", Управління розвитком складних систем, № 38, с. 64–70, 2019. [Видання включено до НБД: BASE; Index Copernicus].*
14. *Oleksandr Terentyev, Svitlana Tsiutsiura, Tetyana Honcharenko, Tamara Lyashchenko, Multidimensional Space Structure for Adaptable in 'International Journal of Recent Technology and Engineering (IJRTE)', ISSN: 2277-3878 (Online), Volume-8 Issue-3, September 2019, Page No. 7753-7758. URL: https://www.ijrte.org/wp-content/uploads/papers/v8i3/C6318098319.pdf*
15. *Гончаренко, Т.А. Метод багатоаспектної класифікації для верифікації багатовимірних інформаційних моделей об'єктів генерального планування [Текст] / Т.А. Гончаренко, В.М. Міхайленко // Управління розвитком складних систем. – 2020. – № 41. – С. 61 – 67; dx.doi.org\10.32347/2412-9933.2020.41.61-67.*

Стаття надійшла до редколегії 10.05.2020

Гончаренко Татьяна Андреевна
Кандидат технических наук, доцент кафедры информационных технологий, *orcid.org/0000-0003-2577-6916*
Киевский национальный университет строительства и архитектуры, Киев

КЛАСТЕРНЫЙ МЕТОД ФОРМИРОВАНИЯ МЕТАДАННЫХ МНОГОМЕРНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ ДЛЯ РЕШЕНИЯ ЗАДАЧ ГЕНЕРАЛЬНОГО ПЛАНИРОВАНИЯ

Аннотация. *Разработан способ формирования метаданных многомерной информационной системы путем сочетания классификационных схем. Каждая классификационная схема представляет собой иерархию значений измерений, относящихся к отдельной структурной компоненты генерального плана (ГП). В основе метода лежит выявление групп значений измерений, связанных с группами значений других измерений. Группы значений различных измерений используются для построения кластеров сочетаний значений измерений. Сочетание кластера формируется декартовым произведением групп значений измерений. Метаданные информационной системы представлены в виде множества допустимых сочетаний значений измерений, которые формируются как набор кластеров. Для решения этой сложной задачи ГП рассматривается как набор структурных компонентов. Из полного набора измерений информационной системы выделяются отдельные наборы измерений, семантически связанные со структурными компонентами ГП. Семантические связи, выявленные в процессе анализа структурной компоненты, позволяющие построить иерархию групп значений измерений и представить их совокупность в виде графа – классификационной схемы, связанной со структурной компонентой. В информационных системах с многоаспектным описанием предметной*

области кубы данных характеризуются большой разреженностью, что затрудняет формирование метаданных. Классификационные схемы описывают отдельные аспекты метаданных, связанные с отдельными структурными компонентами ГП. Сочетание классификационных схем дает возможность получить полное описание метаданных. Использование классификационных схем позволяет разбить задачу описания структуры аналитического пространства многомерной информационной системы на более простые задачи анализа его отдельных структурных компонентов. Сочетание классификационных схем, относящихся к различным структурным компонентам, дает возможность сформировать метаданные информационной системы. В метаданных центральное место занимает множество допустимых сочетаний значений измерений.

Ключевые слова: многомерная информационная система; генеральное планирование; многомерный куб данных; разреженный куб данных; классификационная схема

Honcharenko Tetyana
PhD (Eng.), Associate Professor, Department of Information Technology, *orcid.org/0000-0003-2577-6916*
Kyiv National University of Construction and Architecture, Kyiv

CLUSTER METHOD OF FORMING METADATA OF MULTIDIMENSIONAL INFORMATION SYSTEMS FOR SOLVING GENERAL PLANNING PROBLEMS

Abstract. *A method for generating metadata of a multidimensional information system by combining classification schemes has been developed. Each classification scheme is a hierarchy of measurement values related to a separate structural component of the master plan (GP). The method is based on the identification of groups of measurement values that are associated with groups of values of other dimensions. Groups of values of different dimensions are used to build clusters of combinations of values of measurements. Cluster connections are formed by the Cartesian product of groups of measurement values. The metadata of the information system is presented in the form of a set of valid combinations of measurement values, which are formed as a set of clusters. To solve this complex problem, GP is considered as a set of structural components. From the full set of measurements of the information system, separate sets of measurements are semantically related to the structural components of the GP. The semantic connections revealed in the course of the analysis of a structural component allow to construct hierarchy of groups of values of measurements and to present their set in the form of the graph – the classification scheme connected with a structural component. In information systems with a multifaceted description of the subject area, data cubes are characterized by high sparseness, which complicates the formation of metadata. Classification schemes describe certain aspects of metadata related to certain structural components of GP. The combination of classification schemes makes it possible to obtain a complete description of metadata. The use of classification schemes allows to divide the problem of describing the structure of the analytical space of a multidimensional information system into simpler problems of analysis of its individual structural components. The combination of classification schemes related to different structural components makes it possible to generate information system metadata. In metadata, the central place is occupied by the set of permissible combinations of measurement values.*

Keywords: multidimensional information system; general planning; multidimensional data cube; sparse data cube; classification scheme

References

1. Thomsen, E., (2002). OLAP Solution: Building Multidimensional Information System. NY, Willey Computer Publishing, 688.

2. Mihaylenko, Viktor, Honcharenko, Tetyana, Chupryna, Khrystyna, Andrashko, Yurii, Budnik, Svitlana, (2019). Modeling of Spatial Data on the Construction Site Based on Multidimensional Information Objects. International Journal of Engineering and Advanced Technology (IJEAT), 8, 6, 3934 – 3940. URL: https://www.ijeat.org/wp-content/uploads/papers/v8i6/F9057088619.pdf

3. Hirata, C.M., & Lima, J.C. (2011). Multidimensional cyclic graph approach: representing a data cube without common sub-graphs. Information Sciences, 181, 2626–2655. DOI: 10.1016/j.ins.2010.05.0

4. Salmam, F.Z., Fakir, M., & Errattahi, R., (2016). Prediction on OLAP data cubes. Journal of Information & Knowledge Management, 15, 2, 449 – 458. DOI:10.1142/S0219649216500222

5. Fu, L. (2004). Efficient evaluation of sparse data cubes. Advances in Web-Age Information Management, 336–345. DOI: 10.1007/978-3-540-27772-9_34

6. Romero, O., Pedersen, T.B., Berlanga, R., Nebot, V., Aramburu, M.J., & Simitsis, A., (2015). Using semantic web technologies for exploratory OLAP: A survey. IEEE Transactions on Knowledge and Data Engineering, 27, 2, 571–588. DOI: 10.1109/TKDE.2014.2330822

7. Salmam, F.Z., Fakir, M., & Errattahi, R., (2014). Explanation in OLAP data cubes. Journal of Information Technology Research, 7, 4, 36–78. DOI: 10.4018/jitr.2014100105

8. Orlov, Y., Gaidamaka, Y., Zaripova, E., (2015). Approach to estimation of performance measures for SIP server model with batch arrivals. Distributed Computer and Communication Networks, 601, 141–150. DOI: 10.1007/978-3-319-30843-2_15

9. Viskov, A.V., & Fomin, M.B., (2012). Modeling of analytical measurements in multidimensional databases. Bulletin of Irkutsk State Technical University, 63, 4, 15 – 19.

10. Gomez, L.I., Gomez, S.A., & Vaisman, A.A., (2012). Generic data model and query language for spatiotemporal OLAP cube analysis. Proceedings of the 15-th International Conference on Extending Database Technology EDBT, New York, ACM, pp. 300 – 311.

11. Fomin, M.B., (2017). Description of metadata of multidimensional information systems using the cluster method. Bulletin of the Irkutsk State Technical University, 21, 7, 78 – 86. https:doi.org/10.21285/1814-3520-2017-7-78-86

12. Honcharenko, Tetyana. (2018). The use of BIM-technology to create an information model territories for development. Management of Development of Complex Systems, 33, 131 – 138.

13. Honcharenko, Tetyana, (2019). Object-oriented modeling of spatial objects of general planning. Management of Development of Complex Systems, 38, 64 – 70, dx.doi.org\10.6084/m9.figshare.9788462.

14. Terentyev, Oleksandr, Tsiutsiura, Svitlana, Honcharenko, Tetyana, Lyashchenko, Tamara. (2019). Multidimensional Space Structure for Adaptable. International Journal of Recent Technology and Engineering (IJRTE), 8, 3, 7753 – 7758. URL: https://www.ijrte.org/wp-content/uploads/papers/v8i3/C6318098319.pdf.

15. Honcharenko, Tetyana & Mihaylenko, Victor, (2020). Multi-aspect classification method for verification of multidimensional information models of objects of general planning. Management of Development of Complex Systems, 41, 61 – 67; dx.doi.org\10.32347/2412-9933.2020.41.61-67.

Посилання на публікацію

APA *Honcharenko, Tetyana, (2020). Cluster method of forming metadata of multidimensional information systems for solving general planning problems. Management of Development of Complex Systems, 42, 93 – 101. dx.doi.org\10.32347/2412-9933.2020.42.93-101.*

ДСТУ *Гончаренко, Т.А. Кластерний метод формування метаданих багатовимірних інформаційних систем для розв'язання задач генерального планування [Текст] / Т.А. Гончаренко // Управління розвитком складних систем. – 2020. – № 42. – С. 93 – 101, dx.doi.org\10.32347/2412-9933.2020.42.93-101.*