УДК 004.62:004.492

Метод виділення властивостей, які характеризують об’єкт предметної області

***М. Г. Глава*[1], *Є. В. Малахов*[2]**
[1]Одеський національний політехнічний університет, проспект Шевченка, 1, м. Одеса, 65044, Україна
[2]Одеський національний університет імені І. І. Мечникова, вул. Дворянська, 2, м. Одеса, 65082, Україна
e-mail: glavamg@gmail.com

Запропоновано метод виділення властивостей, які характеризують певний об'єкт предметної області, з метою скорочення трудових і часових витрат на зіставлення об'єктів різних предметних областей при побудові об’єднаної моделі предметної області в процесі інтеграції гетерогенних баз даних. Аналіз властивостей виконується на основі значень статистичних характеристик, отриманих стандартними засобами збору статистичних даних, функціонуючих баз даних. Узгоджений ранг властивостей запропоновано розраховувати за допомогою методу медіанний рангів і медіани Кемені. Для зниження ймовірності помилки, узгодженість значень статистичних характеристик перевіряється на основі коефіцієнта конкордації та перевірки його значущості за допомогою статистики розподілу критерію χ^2 Пірсона.

Ключові слова: *база даних, предметна область, модель предметної області, інформаційна система, властивості об’єктів, статистичні характеристики*

1 Вступ

Однією з актуальних проблем в області інформаційних систем (ІС) є розробка методів інтеграції інформаційних ресурсів, оскільки в процесі розвитку організацій збільшуються потоки даних та поступово ускладнюється архітектура інформаційної системи. Задача ефективного управління інформаційною структурою становиться все гострішою. Крім того, змінюються та/або збільшується кількість проблем, які вирішуються над предметною областю (ПрО), що в свою чергу потребує внесення змін до роботи ІС, тобто інтеграції розрізнених інформаційних потоків, що дозволить отримувати актуальні дані з мінімальними часовими витратами. Існує декілька підходів до інтеграції ІС [1], при цьому одним з перспективних являється дослідження інтеграції даних на семантичному рівні.

2 Постановка задачі

Для побудови будь-якої ІС, в основі якої найчастіше в сучасних бізнес-компаніях лежить реляційна база даних (БД), необхідно спроектувати БД, тобто описати ПрО і побудувати її модель, яка відповідає концептуальній схемі БД [2]. Тобто для інтеграції гетерогенних БД необхідно об’єднати моделі ПрО, а для уникнення надмірності даних виявити однакові об’єкти ПрО та їх властивості.

Класична модель ПрО, яка представлена множиною об’єктів E та зв’язків V між ними, не дозволяє виконати таке порівняння, оскільки одні і ті ж об’єкти можуть мати різні назви, хоч і близькі за семантичним значенням. Семантичний аналіз назв об’єктів достатньо трудовитратний та складний і нетривіальний процес, який не гарантує високої точності такого аналізу. В [3] запропоновано технологію пошуку проекцій одних і тих же об'єктів ПрО, в якій пропонується виділити властивості, які характеризують об’єкт ПрО.

3 Основна частина

Для розв’язання поставленої задачі по аналогії з ідеєю К. Дж. Дейта про узгодження двох моделей даних запропоновано при розгляді моделі об’єкта E ПрО враховувати не тільки його реляційні властивості A, а й екземпляри K, тобто значення властивостей певного об’єкту:

$$E = \langle A, K \rangle \qquad (1)$$

Оскільки кожен об’єкт ПрО, навіть один і той же, може бути описаний різним набором властивостей, запропоновано виділяти дві підмножини властивостей, що характеризують A_c та не характеризують A_u об’єкт ПрО, що дозволить зменшити обсяг даних, що обробляються, за рахунок виключення з аналізу властивостей, які не характеризують цей об’єкт (рисунок 1):

$$A = \langle A_c, A_u \rangle . \qquad (2)$$

Для виділення властивостей, які характеризують певний об’єкт ПрО, запропоновано здійснити збір даних $c_{a_i}^{Ch_n}$ по статистичних характеристиках Ch_n кожної властивості a_i кожного об’єкту ПрО.

Статистичними характеристиками властивостей об’єктів називаються значення, отримані за допомогою стандартних засобів збору статистичних даних за певний період роботи БД за наступними показниками: кількість значень Ct_k, кількість звернень в подіях, які реалізують реляційні операції проекції (select) Ct_s, вибірки (where) Ct_w, з’єднання (join) Ct_{jn}, кількість входжень властивості в тіло тригера або тригерної функції Ct_{tr}, представлення Ct_v та збережені процедури Ct_{pr}.

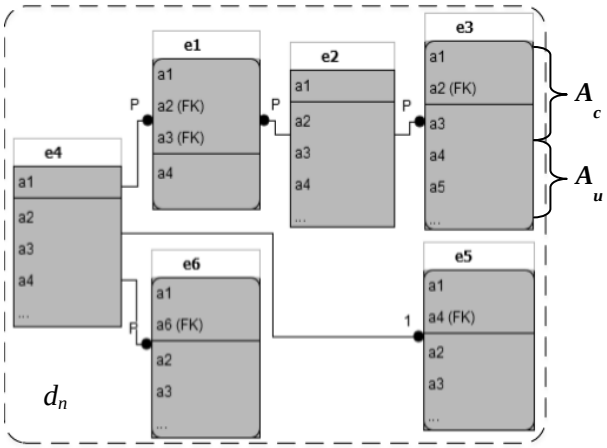

Рисунок 1 *– Виділення властивостей, які характеризують об'єкт e_i ПрО d_n*

Для аналізу статистичних характеристик на першому етапі необхідно розрахувати стандартизовані ранги r_{a_i} властивостей по кожній статистичній характеристиці.

Стандартизація рангів необхідна в тому випадку, якщо по одній і тій же статистичній характеристиці декілька властивостей отримали рівне значення.

Наступним кроком для зменшення вірогідності помилок запропоновано виконати обчислення узгодженого рангу $r_{a_i}^{c}$ кожної властивості об'єкту ПрО двома методами, шляхом розрахунку медіанних рангів $r_{a_i}^{c_M}$ та медіани Кемені $r_{a_i}^{c_K}$.

Метод медіан рангів полягає в тому, що стандартизовані значення, які відповідають одній з статистичних характеристик необхідно розташувати в порядку неубування, тобто перетворити в варіаційний (ранжируваний) ряд [4]. Значення, які стоять на центральних місцях (при непарній кількості статистичних характеристик, як в нашому випадку) або середнє арифметичне двох центральних значень (при парній кількості характеристик) і є медіаною.

Медіана Кемені – окремий випадок визначення емпіричного середнього в просторах нечислової природи [4]. Мінімізація сумарної відстані (метрики) між бінарними відносинами називається «медіаною Кемені» на ім'я її автора Дж. Кемені. Відстань (метрика) між бінарними відносинами A і B, які описуються матрицями a_{ij} та b_{ij}, тобто медіана Кемені дорівнює сумі модулів різниць елементів, що стоять на одних і тих же місцях у відповідних їм матрицях.

Далі пропонується ранжувати властивості за результатами $r_{a_i}^{c_M}$ та $r_{a_i}^{c_K}$ та зіставити їх результати. Впорядкування всієї множини допустимих альтернатив відповідно до їх переваги для оцінювача носить назву ранжування [5]. Номер, який отримують об'єкти оцінки в ході цієї процедури, називають рангом. Це натуральне число, що характеризує порядкове місце оцінюваного об'єкта в групі інших.

Якщо отримані результати ранжування за неубуванням $r_{a_i}^{c_M}$ та $r_{a_i}^{c_K}$ властивості відрізняються, ранг, який має не еквівалентні властивості, приймається за пороговий

$$Z = r_{a_i}^{c}. \quad (3)$$

Якщо результати ранжування запропонованими методами повністю співпадають, Z задається експертом в певній ПрО.

Для відхилення випадкових результатів оцінювання пропонується перевірити узгодженість отриманих значень статистичних характеристик.

Ступінь узгодженості оцінки в теорії рангової кореляції виражається через коефіцієнт конкордації [6] W:

$$W = \frac{12}{N^2(m^3-m) - N\cdot\sum_{j=1}^{N}\sum_{l=1}^{L}\left(t_l^3 - t_l\right)} \cdot \sum_{i=1}^{m}\left(\sum_{n=1}^{N} r_{a_i}^{Ch_n} - \frac{\sum_{i=1}^{m}\sum_{n=1}^{N} r_{a_i}^{Ch_n}}{m}\right)^2 \quad (4)$$

Коефіцієнт конкордації може приймати значення від 0 до 1, причому 0 означає повну неузгодженість оцінки, а 1 – повне узгодження оцінки.

Перед прийняттям рішення пропонується перевірити значимість коефіцієнта конкордації за допомогою статистики розподілу критерія χ^2 Пірсона.

За допомогою критерію χ^2 можна порівнювати вибірки, які мають альтернативні ознаки, а також оцінювати вірогідність кореляції між альтернативними ознаками [7]. Як і інші критерії згоди (Колмогорова А, Романовского, Фішера Б, Ястремского), χ^2 являє собою деяку величину, яка оцінюється з певною ймовірністю. Він може приймати різні завжди додатні значення (малі й великі). При χ^2=0 слід вважати, що відміни між частотами порівнюваних рядів розподілу відсутні. Даний критерій не рекомендується використовувати для оцінки малих вибірок.

Якщо $\chi^2_{розраховане} > \chi^2_{табличне}$, гіпотеза про значимість нулю коефіцієнта конкордації відхиляється, а отже узгодженість отриманих значень статистичних характеристик присутня.

Останнім кроком виділяємо підмножини властивостей A_c та A_u:

$$\forall a_i = \begin{cases} a_i \in A_c \,|\, r_{a_i}^{c} \le Z \\ a_i \in A_u \,|\, r_{a_i}^{c} > Z \end{cases}. \quad (5)$$

Властивості A_u, що не характеризують об'єкт ПрО, виключаються з аналізу. Але, якщо ступінь подібності об'єктів буде низьким, тобто виникне необхідність уточнення чи дійсно об'єкти різні, буде проведений аналіз і зіставлення властивостей A_u.

4 Апробація запропонованого методу

Перевіримо працездатність запропонованого методу на реальних даних. Нехай існує об'єкт e_i ПрО з 5 властивостями *a*.

Приклад зібраних даних $c_{a_i}^{Ch_n}$ по статистичних характеристиках Ch_n по властивостям об'єкту e_i представлено на рисунку 2.

Ch_n \ a_i	a_1	a_2	a_3	a_4	a_5
Ct_k	$c_{a_1}^{Ct_k}=200$	$c_{a_2}^{Ct_k}=200$	$c_{a_3}^{Ct_k}=198$	$c_{a_4}^{Ct_k}=197$	$c_{a_5}^{Ct_k}=195$
Ct_s	$c_{a_1}^{Ct_s}=1000$	$c_{a_2}^{Ct_s}=1020$	$c_{a_3}^{Ct_s}=900$	$c_{a_4}^{Ct_s}=851$	$c_{a_5}^{Ct_s}=300$
Ct_w	$c_{a_1}^{Ct_w}=50$	$c_{a_2}^{Ct_w}=64$	$c_{a_3}^{Ct_w}=16$	$c_{a_4}^{Ct_w}=0$	$c_{a_5}^{Ct_w}=20$
Ct_{jn}	$c_{a_1}^{Ct_{jn}}=30$	$c_{a_2}^{Ct_{jn}}=25$	$c_{a_3}^{Ct_{jn}}=10$	$c_{a_4}^{Ct_{jn}}=20$	$c_{a_5}^{Ct_{jn}}=20$
Ct_{tr}	$c_{a_1}^{Ct_{tr}}=3$	$c_{a_2}^{Ct_{tr}}=5$	$c_{a_3}^{Ct_{tr}}=2$	$c_{a_4}^{Ct_{tr}}=4$	$c_{a_5}^{Ct_{tr}}=1$
Ct_v	$c_{a_1}^{Ct_v}=2$	$c_{a_1}^{Ct_v}=2$	$c_{a_1}^{Ct_v}=2$	$c_{a_4}^{Ct_v}=2$	$c_{a_5}^{Ct_v}=0$
Ct_{pr}	$c_{a_1}^{Ct_{pr}}=3$	$c_{a_2}^{Ct_{pr}}=2$	$c_{a_3}^{Ct_{pr}}=3$	$c_{a_4}^{Ct_{pr}}=3$	$c_{a_5}^{Ct_{pr}}=1$

Рисунок 2 *– Статистичні оцінки властивостей об'єкта*

Переведемо отриманні значення статистичних оцінок у рангову шкалу та стандартизуємо їх, оскільки серед значень по характеристиках є рівні. Стандартизовані ранги представлено на рисунку 3.

Розрахуємо узгоджені ранги кожної властивості методом медіан $r_{a_i}^{c_M}$ та медіани Кемені $r_{a_i}^{c_K}$ (рисунок 3).

Оскільки властивості a_3 та a_4 мають однаковий узгоджений ранг $r_{a_i}^{c_M}$, розрахований методом медіан, необхідно стандартизувати його:

$$\frac{3+4}{2}=3{,}5\,.$$

Проранжуємо властивості за результатами $r_{a_i}^{c_M}$ (рисунок 4, *а*) та $r_{a_i}^{c_K}$ (рисунок 4, *б*).

Як видно з рисунку 4 результати ранжування властивостей розходяться. За результатами розрахунку узгодженого рангу методом медіанних рангів властивості a_3 та a_4 мають еквівалентний ранг.

Ch_n \ a_i	a_1	a_2	a_3	a_4	a_5
Ct_k	1,5	1,5	3	4	5
Ct_s	2	1	3	4	5
Ct_w	1	2	4	5	3
Ct_{jn}	3	3	3	3	3
Ct_{tr}	3	1,5	4	1,5	5
Ct_v	3	1	3	3	5
Ct_{pr}	4	2	4	4	1
$r_{a_i}^{c_M}$	2	1	4	4	5
$r_{a_i}^{c_K}$	2	1	3	4	5

Рисунок 3 *– Стандартизовані та узгоджені ранги властивостей об'єкта*

Ch_n \ a_i	a_2	a_1	a_3	a_4	a_5
Ct_k	1,5	1,5	3	4	5
Ct_s	1	2	3	4	5
Ct_w	2	1	4	5	3
Ct_{jn}	3	3	3	3	3
Ct_{tr}	1,5	3	4	1,5	5
Ct_v	1	3	3	3	5
Ct_{pr}	2	4	4	4	1
$r_{a_i}^{c_M}$	1	2	3,5		5

а)

Ch_n \ a_i	a_2	a_1	a_3	a_4	a_5
Ct_k	1,5	1,5	3	4	5
Ct_s	1	2	3	4	5
Ct_w	2	1	4	5	3
Ct_{jn}	3	3	3	3	3
Ct_{tr}	1,5	3	4	1,5	5
Ct_v	1	3	3	3	5
Ct_{pr}	2	4	4	4	1
$r_{a_i}^{c_K}$	1	2	3	4	5

б)

Рисунок 4 *– Ранжування властивостей шляхом розрахунку медіанних рангів (а) та медіани Кемені (б)*

А за результатами розрахунку на підставі медіани Кемені властивість a_3 має перевагу перед властивістю a_4. Отже, згідно (3)

$$Z = 3.$$

Розрахуємо згідно (4) ступінь узгодженості отриманих значень по статистичних характеристиках для переконання, що дані значення не отримано випадково

$$W = 0{,}59.$$

Коефіцієнт конкордації W значно більше 0, отже, узгодженість статистичних характеристик присутня.

Перевіримо значимість коефіцієнта конкордації за допомогою статистики розподілу критерія χ^2 Пірсона:

$\chi^2_{розраховане}$ = 11,88; $\chi^2_{табличне}$ = 9,48 при 5% рівні коефіцієнта значимості.

$\chi^2_{розраховане} > \chi^2_{табличне}$, гіпотеза про значимість нулю коефіцієнта конкордації відхиляється, а отже узгодженість отриманих значень статистичних характеристик присутня.

5 Висновки

Експериментальний аналіз показав, що за рахунок запропонованого методу виділення властивостей, які характеризують об'єкт ПрО, кількість властивостей для

подальшої обробки було скорочено в даному випадку на 40%. Експериментальні дослідження на інших даних показали зменшення властивостей на 10-32% в залежності від об'єкту за рахунок виключення з аналізу властивостей, які не характеризують цей об'єкт.

Література

1. **Васильева Т. П.** Основные проблемы и методы интеграции баз данных / Т. П. Васильева, М. Г. Глава. // Первый независимый научный вестник. – 2015. – №1. – С. 28–32.
2. **Glava M.** Information Systems Reengineering Approach Based on the Model of Information Systems Domains / Maria Glava, Valery Malakhov // International Journal of Software Engineering and Computer Systems (IJSECS).– University Malaysia Pahang.–2018.– Vol. 4.– P. 95–105.
3. **Glava M.** Searching Similar Entities in Models of Various Subject Domains Based on the Analysis of Their Tuples / Maria Glava, Eugene Malakhov // 2016 International Conference on Electronics and Information Technology (EIT), May 23–27, 2016, Odesa, Ukraine, 2016.– P. 97–100. (DOI: 10.1109/ICEAIT.2016.7501001; EID: 2-s2.0-84979503116).
4. **Коваленко И. И.** Методы экспертного оценивания сценариев: учебное пособие [Электронный ресурс] / И.И. Коваленко, А.В. Швед.– Николаев: Изд-во ЧГУ им. Петра Могилы, 2012.– 156 с.– Режим доступу: *http://lib.chdu.edu.ua/index.php?m=2&b=308*
5. **Писарева О. М.** Методы социально-экономического прогнозирования: Учебник / О. М. Писарева.– ГУУ – НФПК.– М., 2003.
6. **Панов В.С.** Автоматизация расчета коэффициента конкордации и выявления согласованности мнений экспертов [Электронный ресурс] / В. С. Панов, А. Е. Суслов.– Режим доступа: *http://conf.sfukras.ru/sites/mn2011/thesis/s3/s3_081.pdf*
7. Основні аспекти і умови застосування Хі-квадрат критерію [Електронний ресурс].– Режим доступу: *http://pidruchniki.com/14940807/statistika/osnovni_aspekti_umovi_zastosuvannya_kvadrat_kriteriyu*

Отримана в редакції 12.02.2018, прийнята до друку 06.03.2018

Method of isolating the properties that characterize the domain object

M. G. Glava, E. V. Malakhov
[1]Odessa National Polytechnic University, 1, Shevchenko Ave., Odessa, 65044, Ukraine
[2]Odessa I. I. Mechnikov National University, str. Dvoryanskaya 2, Odessa, 65026, Ukraine
e-mail: glavamg@gmail.com

Modern information systems are a set of information which is contained in the databases and information technologies providing its processing. In enterprise information systems there is a problem of interaction of the databases realized in different DBMSs. There is urgent problem of association of the existing heterogeneous databases integrated into a common information space, i.e. association of information models of subject domains. Method is proposed for isolating the properties characterizing a certain object of the subject domain in order to reduce labor and time costs for comparing objects of different subject domains when constructing a general subject domain model in the process of integrating heterogeneous databases. The analysis of properties based on the values of statistical characteristics which received of the standard means of collecting statistical data. The agreed rank of properties is proposed to be calculated using the method of median ranks and the median of Kemeni. Match the ranking result, based on integral ranks. Select a value threshold: if the ranking results diverge, the threshold is equal to the rank on which it does not follow; if the ranking results coincide, the rank is set by the expert in a particular subject domain. To reduce the probability of error, the consistency of the values of the statistical characteristics is verified on the basis of the concordance coefficient and verification of its significance using the Pearson χ^2 criterion distribution statistics. The properties that do not characterize the object of the subject domain, are excluded from the analysis. But, if the degree of similarity of objects will be low, it will be necessary to clarify whether the objects are different.

Keywords: *database, subject domain, model of subject domain, information system, properties of objects, statistical characteristics*

References

1. **Vasilyeva T.P., Glava M.G.** (2015) Osnovnye problemy i metody integratsii baz dannykh. *Pervy nezavisimy nauchny vestnik*, No 1, pp. 28–32.
2. **Glava M., Malakhov V.** (2018) Information Systems Reengineering Approach Based on the Model of Information Systems Domains. *International Journal of Software Engineering and Computer Systems (IJSECS)*, University Malaysia Pahang, vol. 4, pp. 95–105.
3. **Glava M., Malakhov E.** (2016) Searching Similar Entities in Models of Various Subject Domains Based on the Analysis of Their Tuples, *2016 International Conference on Electronics and Information Technology (EIT), Odesa, Ukraine*, 2016, pp. 97–100. (DOI: 10.1109/ICEAIT.2016.7501001; EID: 2-s2.0-84979503116).
4. **Kovalenko I. I., Shved A. V.** (2012) Metody ekspertnogo otsenivaniya stsenariyev: uchebnoye posobiye, Nikolayev: ChGU im. Petra Mogily, p. 156. URL: http://lib.chdu.edu.ua/index.php?m=2&b=308
5. **Pisareva O. M.** (2003) Metody sotsialno-ekonomicheskogo prognozirovaniya: Uchebnik, GUU - NFPK, Moskva.

6. **Panov V. S., Suslov A. Ye.** Avtomatizatsiya rascheta koeffitsiyenta konkordatsii i vyyavleniya soglaso-vannosti mneny ekspertov. URL: http://conf.sfu-kras.ru/sites/mn2011/thesis/s3/s3_081.pdf
7. Osnovni aspekty i umovy zastosuvannia Xi-kvadrat kryteriiu. URL: http://pidruchniki.com/14940807/statistika/osnovni_aspekti_umovi_zastosuvannya_kvadrat_kriteriyu

Received 12 February 2018
Approved 06 March 2018
Available in Internet 30 April 2018