УДК 004.658.2

БУЙ Д.Б., ПУЗІКОВА А.В.

ОГЛЯД ТЕОРІЇ НОРМАЛІЗАЦІЇ В РЕЛЯЦІЙНИХ БАЗАХ ДАНИХ

Вперше термін «нормалізація» застосував у 1970 р. Е. Кодд для назви процедури усунення непростих доменів [1]. Інтерпретація першої нормальної форми (1НФ) змінювалась в залежності від інтерпретації поняття «непростого домену» [2–4]. Можливість використовувати в якості елементів домену групи (groups) та відношення відстоювали автори робіт [5, 6].

У 1971 р. Е. Кодд у роботі [7] вказує на надлишковість даних та аномалії, які виникають при здійсненні операцій над відношеннями, вперше представляє концепцію функціональної залежності (ФЗ), демонструє можливість її використання для розв'язання проблем проектування баз даних (БД), наводить означення другої нормальної форми (2НФ), транзитивної ФЗ та третьої нормальної форми (3НФ). У науковій літературі зустрічаються їх різні інтерпретації. Зокрема, відмінності в означеннях 3НФ пояснюються різними уточненнями, які накладаються на визначення транзитивної залежності [8, 9].

Відкриття аксіом та правил виведення ФЗ із заданої множини ФЗ [10] та побудова аксіоматики Армстронга для ФЗ [11] дали можливість розробити алгоритми обчислення так званого канонічного покриття (за термінологією Мейера [9]) для заданої множини ФЗ та замикання для множини атрибутів [9, 12–14].

Одним із способів приведення відношення до 3НФ є декомпозиція, обґрунтуванням якої стала теорема Хеза (Heath) [15]. До «підводних каменів» декомпозиції належить залежність проекцій (за термінологією Ріссанена (Rissanen) [16]), яка може стати причиною аномалій. Дана проблема обговорюється у статті А. Філліповича [17], який розглянув взаємні ФЗ (ВФЗ), дослідив їх властивості, побудував алгоритм виявлення ВФЗ, запропонував поняття взаємно-незалежної нормальної форми, яку можна розглядати як синонім ациклічної БД, та спосіб зведення до неї.

Інший спосіб запропонував Бернштейн (Bernstein), який побудував алгоритм синтезу повної схеми БД у 3НФ для заданої множини ФЗ [18].

Специфіка різних підходів до задачі проектування схеми реляційної БД викликана відмінностями у формальних визначеннях еквівалентності та критеріїв якості схеми [19].

Відомими класичними алгоритмами зведення схеми відношення до 3НФ є алгоритми Ульмана [13], Делобеля-Гейсі (Delobel-Gasey) [10], результатами яких не завжди є схема у 3НФ, Берштейна [18], Іслура (Isloor) [20], Неклюдової-Цаленка [21], який дає кількісно оптимальну схему БД, Мейера, реалізований через побудову кільцевих покриттів [9]. Переваги та недоліки більшості з вказаних алгоритмів, а також їх відповідність різним визначенням еквівалентності реляційних схем розглянуті у монографії [8]. Пошук ефективних алгоритмів розв'язання задачі синтезу оптимальної схеми БД у 3НФ продовжується і сьогодні [14, 22, 23].

Недоліки 3НФ були враховані у роботі Хеза [15] при формулюванні означення посиленої 3НФ та, пізніше, у роботі Кодда [24] (інша назва означення – нормальна форма Бойса-Кодда (НФБК)). Одним з перших відомих алгоритмів зведення «майже» до НФБК є алгоритм Берштейна [18], який дозволяє усувати транзитивні залежності первинних атрибутів від ключів, що не містять ці атрибути. Більш пізні алгоритми наводяться, наприклад, у роботах [25, 26].

З введенням у розгляд багатозначних залежностей (БЗЗ) [27] Р. Фагін (Fagin) досліджує їх властивості та визначає нову четверту нормальну форму (4НФ) [28]. Строгий та повний набір правил виведення для БЗЗ, а також правила, які пов'язують ФЗ та БЗЗ, представлені у статті [29], незалежність побудованої системи аксіом обговорюється у роботі Мендельзона (Mendelzon) [30], а її повнота – у статті Біскапа (Biskap) [31].

Залежності з'єднання (ЗЗ) та аномалії, які вони викликають, були розглянуті Ріссаненом (Rissanen) [32] та досліджені у роботах [33, 34]. Концепція п'ятої нормальної форми (5НФ) представлена Р. Фагіним [35]. З результату про відсутність повної скінченної множини правил

виведення для ЗЗ, який наводиться у роботі С. Петрова [36], випливає неможливість побудови повної та коректної аксіоматики, звідси – неможливість побудови канонічного покриття.

Викладенню результатів з теорії залежностей та теорії нормалізації для реляційних БД присвячені монографії Д. Мейера [9], В. Дрібаса [8], М. Цаленка [37].

Для класичних 1-5НФ пропонувались різноманітні варіанти їх покращення, наприклад, покращена 3НФ (An Improved Third Normal Form) [38], а також – введення інших видів нормальних форм, наприклад, нормальної форми з елементарним ключем (Elementary Key Normal Form), яка займає проміжне положення між 3НФ та НФБК [39], нормальної форми з повним ключем (Key-Complete Normal Form) [40], кортеже-необхідної нормальної форми (Essential Tuple Normal Form) [41], яка визначається в термінах ФЗ и ЗЗ та займає проміжне положення між 4НФ і 5НФ, доменно-ключової нормальної форми (ДКНФ) [42].

Зокрема, концепція ДКНФ базується на поняттях залежності ключа та залежності домена: «Відношення знаходиться у ДКНФ тоді і тільки тоді, коли кожне обмеження з його схеми є логічним наслідком з об'єднання множин залежності ключа та залежності домена». Але під обмеженнями в означенні ДКНФ розуміються не тільки ФЗ, БЗЗ та ЗЗ, означення яких використовуються у попередніх нормальних формах, а й усі, які можна записати у вигляді висловлювання логіки предикатів 1-го порядку. У роботі [43] пропонується форма запису для усіх відомих на той час залежностей:

$$(\forall x_1 ... x_m)((A_1 \wedge ... \wedge A_n) \Rightarrow \exists y_1 ... y_r (B_1 ... B_s)), \quad (1)$$

де A_i – атомарна реляційна формула для представлення входження індивідних змінних $z_1 ... z_d$ у d-арне відношення P вигляду $P_{z_1 ... z_d}$; B_i – або атомарна реляційна формула, або рівність $x = y$, де x і y – індивідні змінні. Зокрема, вбудовані БЗЗ (ВБЗЗ) можна задати у вигляді

$$(\forall ab_1 b_2 c_1 c_2 d_1 d_2)((Pab_1 c_1 d_1 \wedge Pab_2 c_2 d_2) \Rightarrow \exists d_3 Pab_1 c_2 d_3). \quad (2)$$

Оскільки задача перевірки імплікації для таких видів обмежень як ВБЗЗ є нерозв'язною [37, 43], то неможливо побудувати алгоритм синтезу реляційної схеми у ДКНФ. До такого ж висновку можна прийти, врахувавши включення 5НФ у ДКНФ за умови нескінченності доменів та нерозв'язності задачі синтезу оптимальної схеми у 5НФ.

Окремим видом є шоста нормальна форма (6НФ), яка була введена у 2002 р. для хронологічних БД [44]. Одним з сучасних напрямків розвитку класичної теорії нормалізації в реляційних БД є поширення її принципів на нечіткі реляційні БД [45, 46].

Підведемо підсумки. Незважаючи на вагомі результати, теорія нормалізації в реляційних БД носить фрагментарний характер і далека ще до задовільного завершення. Автори статті також мають певні результати з даної тематики: було запропоновано строге математичне доведення коректності та повноти аксіоматики Армстронга, виконане в традиціях математичної логіки шляхом встановлення збіжності семантичного та синтаксичного слідувань [47]; запропоновано критерій повноти аксіоматики Армстронга в термінах потужностей множини атрибутів та універсального домену [48]. Зауважимо, що нормалізація, основною метою якої є підтримка цілісності даних, вступає в суперечність з ефективністю опрацювання операцій в БД; саме тому зараз вже можна говорити про початок створення теорії денормалізації та про природнє (точніше кажучи, оптимальне за певним критерієм) поєднання нормалізації та денормалізації.

1. ЛІТЕРАТУРА

2. Codd, E. F. A Relational Model of Data for Large Shared Data Banks / E. F. Codd // Communications of the ACM. – 1970. – Vol. 13, № 6. – P. 377–387.
3. Дейт, К. Дж. Введение в системы баз данных / К. Дж. Дейт. – М.: Наука, 1980. – 464 с.
4. Elmasri, R. Fundamentals of Database Systems / R. Elmasri, S. R. Navathe. – Massachusetts: Addison-Wesley, 2000. – 893 p.
5. Дейт, К. Дж. Введение в системы баз данных / К. Дж. Дейт. – М.: Вильямс, 2005. – 1328 с.
6. Jaeschke, G. Remarks on the Algebra of Non First Normal Form Relations / G. Jaeschke, H. J. Schek // Proc. of the 1st ACM SIGACT-SIGMOD symposium on Principles of database systems, Los Angeles, California, 1982. – P. 124-138.

7. Makinouchi, A. A Consideration of Normal Form on Not-necessarily Normalized Relations in the Relational Data Model / A. Makinouchi // Proc. of the Third Intern. Conf. on Very Large Data Bases, October 6-8, 1977, Tokyo, Japan. IEEE Computer Society. – 1977. – P. 447-453.
8. Codd, E.F. Further Normalization of the Data Base Relational Model / E.F. Codd // IBM Research Report RJ909 (August 31, 1971). Republished in Randall J. Rustin (ed.), Data Base Systems: Courant Computer Science Symposia Series 6. Prentice-Hall, 1972.
9. Дрибас, В. П. Реляционные модели баз данных / В. П. Дрибас. – Минск: БГУ, 1982. – 192 с.
10. Мейер, Д. Теория реляционных баз данных / Д. Мейер. – Москва: Мир, 1987. – 608 с.
11. Delobel, C. Decomposition of a database and the theory of Boolean switching function / C. Delobel, R. Gasey // IBM Journal of Research and Development. – 1973. – Vol. 17, № 5. – P. 374-386.
12. Armstrong, W. W. Dependency structures of data base relationships / W. W. Armstrong // Proc. IFIP ‘74, North Holland Pub. Co., Amsterdam, 1974. – P. 580-583.
13. Beeri, C. Computational problems related to the design of normal form relation schemas / C. Beeri, P. A. Bernstein // ACM Transactions on Database Systems. – 1979. – Vol. 4, № 1. – P. 30-59.
14. Ульман, Дж. Основы систем баз данных / Дж. Ульман. – М.: Фин. и стат., 1983. – 334 с.
15. Григорьев, Ю. А. Алгоритм синтеза частично оптимальной схемы реляционной базы данных / Ю. А. Григорьев // Электронное научно-техническое издание «Наука и образование», 2012. – № 1. – Режим доступа: http://technomag.edu.ru/doc/294486.html.
16. Heath, I. J. Unacceptable File Operations in Relational Database / I. J. Heath // ACM SIGFIDET Workshop on Data Description, Access, and Control. – San Diego. – 1971. – P. 19-33.
17. Rissanen, J. Independent components of relations / J. Rissanen // ACM Transactions on Database Systems. – 1977. – Vol. 2, № 4. – P. 317-325.
18. Филлипович, А. Взаимные функциональные зависимости / А. Филлипович // Системный администратор. – 2002. – № 1. – С. 84-89.
19. Bernstein, P. A. Synthesizing Third Normal Form relations from functional dependencies / P. A. Bernstein // ACM Transactions on Database Systems. – 1976. – Vol. 1, № 4. – P. 277-298.
20. Beeri, C. A sophisticate’s introduction to database normalization theory / C. Beeri, P. Bernstein, N. Goodman // Proceedings of 4th International Conference on Very Large Data Bases, West Berlin, 1978. – P. 113-124.
21. Isloor, S. S. An algorithm with logical simplicity for designing third normal form relations data base schema for functional dependencies / S. S. Isloor // Proceedings of International Conference on DBMS (ICMOD 78), Fast Milano, Italy, 1978. – P. 31-50.
22. Неклюдова, Е. А. Синтез логической схемы реляционной базы данных / Е. А. Неклюдова, М. Ш. Цаленко // Программирование. – 1979, № 6. – С. 58-68.
23. Зорин, И. Теоретико-графовое приведение реляционной базы данных к третьей нормальной форме Э. Кодда / И. Зорин // Электронное научное издание «Устойчивое инновационное развитие: проектирование и управление», 2009. – Т. 5. – С. 50-59.
24. Виноградова, М. В. Конструктор баз данных на основе сущностей и их реквизитов с возможностью нормализации / М. В. Виноградова, Э. Г. Игушев // Электронное научно-техническое издание «Наука и образование», 2011. – № 10.
25. Codd, E. F. Recent Investigations into Relational Data Base Systems / E. F. Codd // Proceedings of IFIP Congress 74, Stockholm, August 5-10, 1974. North-Holland, 1974. – P. 1017-1021.
26. Lin, W. Y. Efficient algorithm for BCNF-decomposition / W.-Y. Lin // Information and Software Technology. – 1992. – Vol. 34, № 5. – P. 308-312.
27. Bahmani, A. Automatic database normalization and primary key generation / A. Bahmani, M. Naghibzadeh, B. Bahmani // CCECE/CCGEI May 5-7 2008, Niagara Falls. – P. 11-16.
28. Zaniolo, C. Analysis and design of relational schemata for database systems: Ph.D. dissertation, Tech. Rep. UCLA-Eng-7769, Dep. Computer Science, Univ. Calif. at Los Angeles, July 1976.
29. Fagin, R. Multivalued Dependencies and a New Normal Form for Relational Databases / R. Fagin // ACM Transactions on Database Systems. – 1977. – Vol. 2, № 1. – P. 262-278.

30. Beeri, C. A complete axiomatization for functional and multivalued dependencies / C. Beeri, R. Fagin, J. Howard // Proc. ACM-SIGMOD Conf. (Toronto, Canada, Aug. 3-5) ACM, New York, 1977. – P. 47-61.
31. Mendelzon A. On axiomatizing multivalued dependencies in relational databases / A. Mendelzon // ACM Transactions on Database Systems. – 1979. – Vol. 26, № 1. – P. 37-44.
32. Biskup, J. Inferences of multivalued dependencies in fixed and undetermined universes / J. Biskup // Theoretical Computer Science. – 1980. – Vol. 10, № 1. – P. 93-105.
33. Rissanen, J. Independent components of relations / J. Rissanen // ACM Transactions on Database Systems. –1977. – Vol. 2, № 4. – P. 317-325.
34. Aho, A. V. The theory of joins in relational databases / A. V. Aho, C. Beeri, J. D. Ullman // Proc. 18th Symp. on Foundations of Computer Science, Providence, R.I., 1977. – P. 107-113.
35. Dayal, U. The fragmentation problem: lossless decomposition of relations into files / U. Dayal, P. A. Bernstein // Proceedings of the ACM SIGMOD international conference on Management of data, 1979. – P. 143-151.
36. Fagin, R. Normal Forms and Relational Database Operators / R. Fagin // Proceedings of the ACM SIGMOD International Conference on Management of Data (Boston, Mass., May 30-June 1), ACM, New York, 1979. – P. 153-160.
37. Петров, С. В. Об аксиоматизации зависимостей по соединению / С. В. Петров // Применение методов математической логики: Тезисы докл. IV Всес. конф. «Представление знаний и синтез программ». – Таллин: АН ЭССР, 1986. – С. 151-152.
38. Цаленко, М. Ш. Моделирование семантики в базах данных / М. Ш. Цаленко. – М.: Наука, 1989. – 287 с.
39. Vincent, M.W. Redundancy Elimination and a New Normal Form for Relational Database Design / M. W. Vincent // In Semantics in Databases, vol. 1358 of LNCS. – 1998. – P. 247-264.
40. Darwen, H. A Normal Form for Preventing Redundant Tuples in Relational Databases / H. Darwen, C. Date, R. Fagin // Proceedings of the 15th International Conference on Database Theory – ICDT'2012, March 26–30, 2012, Berlin, Germany. – P. 114-126.
41. Fagin, R. A Normal Form for Relational Databases That Is Based on Domains and Keys / R. Fagin // Communications of the ACM. – 1981. – Vol. 6. – P. 387-415.
42. Fagin, R. The theory of data dependencies – a survey / R. Fagin, M. Y. Vardi // In Mathematics of Information Processing, Proc. Symposia in Applied Mathematics – vol. 34, American Mathematical Society, 1986. – P. 19-71.
43. Date, C. J. Temporal Data and the Relational Model / C. J. Date, H. Darwen, N. Lorentzos. – Morgan Kaufmann, 2002. – 422 p.
44. Chen, G. Normalization based on ffd in a fuzzy relational data model / G. Chen, E. E. Kerre, J. Vandenbulcke // Inform Syst, 1996. – Vol. 21. – P. 299-310.
45. Bahar, Ö. Normalization and Lossless Join Decomposition of Similarity-Based Fuzzy Relational Databases / Ö. Bahar, A. Yazıcı // International Journal of Intelligent Systems, 2004. – Vol. 19. – P. 885-917. Published online in Wiley InterScience (www.interscience.wiley.com).
46. Буй, Д. Б. Повнота аксіоматики Армстронга / Д. Б. Буй, А. В. Пузікова // Вісник КНУ ім. Т. Шевченка: Фіз.-мат. науки, № 3, 2011. – С. 103-108.
47. Буй, Д. Б. Критерій повноти аксіоматики Армстронга / Д. Б. Буй, А. В. Пузікова // Матер. міжн. конф. «Теоретичні та прикладні аспекти побудови програмних систем» – TAAPSD'2011 (Ялта, 19-23 вересня 2011 року). – С. 30-34.

БУЙ Дмитро Борисович – д.ф.-м.н., професор, професор кафедри теорії і технології програмування факультету кібернетики КНУ ім. Т. Шевченка.

Наукові інтереси: *теорія баз даних, теорія програмних алгебр композиційного типу, композиційна семантика мови SQL, теорія нерухомих точок, сучасні CASE-засоби.*

ПУЗІКОВА Анна Валентинівна – аспірантка кафедри теорії і технології програмування факультету кібернетики КНУ ім. Т. Шевченка.

Наукові інтереси: *нормалізація в реляційних базах даних.*