УДК 004.652

# ЗАСТОСУВАННЯ ФАКТОЛОГІЧНОЇ РЕЛЯЦІЙНОЇ МОДЕЛІ ДЛЯ ІНТЕГРАЦІЇ НЕОДНОРІДНИХ СТРУКТУР ДАНИХ

**А.Ю. Берко**
Кандидат технічних наук, доцент
Кафедра інформаційних систем та мереж
Національний університет "Львівська політехніка"
вул. С. Бандери, 12, м. Львів, Україна, 79013
Контактний тел.: 8 (032) 258-25-38
E-mail: BerkoAndriy@Ya.Ru

***Описано деякі можливості подання неоднорідних та слабко структурованих даних. Запропоновано фактологічну реляційну структуру даних для вирішення поставленої задачі***

***Ключові слова: бази даних, інтеграція даних, реляційна модель, фактологічна структура даних, неоднорідні дані***

***Рассмотрены определенные возможности представления неоднородных и слабо структурированных данных. Предложена фактологическая структура данных для решения поставленной задачи***

***Ключевые слова: базы данных, интеграция данных, реляционная модель, фактологическая структура данных, неоднородные данные***

***Any possibilities of presentation of heterogeneous and semi structured data in relation form are considered. Factological relation structure of data has been proposed for solution of this problem***

***Key words: data bases, data integration, relational data model, factological data structure, heterogeneous data***

## Вступ

Актуальність проблеми інтеграції ресурсів інформаційних систем обумовлена, насамперед, такими факторами. По-перше – постійне інтенсивне зростання обсягів даних, які є об'єктом застосування інформаційних технологій у вирішенні задач різноманітних сфер людської діяльності. По-друге – велика кількість і різноманіття способів, форм і форматів подання даних, а також методів та засобів їх опрацювання. По-третє – активний розвиток інформаційних ресурсів суспільного використання, таких як сховища даних, корпоративні системи, електронні бібліотеки, інформаційні web-системи, системи типу "cloud computing" тощо.

Загалом, об'єктом інтеграції можуть бути різні ресурси інформаційних систем. Розрізняють такі напрями інтеграції як інтеграція застосувань, інтеграція платформ, інтеграція процесів та інтеграція даних. У запропонованій роботі розглянуто один із підходів до інтегрованого подання різнорідних даних, який ґрунтується на застосуванні принципів реляційної моделі, слабкоструктурованих даних та поняття факту, як одиниці зображення та сприйняття даних.

## 1. Проблеми та методи інтеграції даних

Основною метою інтеграції даних є формування глобального інформаційного ресурсу на основі множини локальних ресурсів для спільного узгодженого застосування. Інтеграція не є простим механічних об'єднанням даних отриманих з різних джерел. Процедури інтеграції передбачають вирішення цілої низки завдань відбору, перетворення, узгодження, об'єднання, контролю якості даних та багатьох інших [3].

Задачі спільного опрацювання даних різної природи та формату сьогодні вирішують у багатьох сферах застосування інформаційних технологій – корпоративних системах, системах комп'ютерного моніторингу, в електронному бізнесі, у системах прийняття рішень та бізнес-аналітики тощо.

Сьогодні проблемам інтеграції даних приділяється особлива увага як з боку провідних виробників систем і засобів управління базами даних (Microsoft, Oracle, IBM, SAS, SAP, Informatica) так і в середовищі міжнародних некорпоративних структур таких як W3C, OASIS, IntegrationConsortium та інших.

Найсуттєвішими проблемами в галузі інтеграції даних є велика кількість різноманітних підходів і технологій, які часто є несумісними між собою та відсутність єдиної теоретичної моделі та методики інтеграції даних, незалежної від їх змісту, формату, засобів реалізації та призначення. Загалом ситуація в сфері інтеграції даних є подібною до ситуації в галузі баз даних до запровадження реляційної моделі.

Окремою проблемою є інтеграція структурованих (баз даних) з, так званими, слабкоструктурованими чи напівструктурованими інформаційними ресурсами.

Якщо проблема створення єдиного середовища опрацювання баз даних є, загалом, достатньо добре дослідженою і забезпеченою відповідними методиками та технологіями [3], то спільне застосування та опрацювання даних неоднорідної структури сьогодні продовжує залишатися проблемним. Особливістю неоднорідних структур є різноманітність їх форми та змісту, способів і засобів подання та опрацювання, а також, дуже часто, неповнота, неточність і часткова невизначеність. Використання традиційних технологій баз даних у таких застосуваннях не завжди є ефективним, а часто неприйнятним, а опрацювання структурованих баз даних за принципами слабкоструктурованих призводить до втрати значної частки їх властивостей.

В цьому напрямі, на думку автора, найпродуктивнішим є підхід, який поєднує функціональні і технологічні можливості та теоретичний апарат баз даних з вільним форматом та широтою спектру подання і застосування слабко-структурованих даних.

## 2. Сучасні підходи до реляційної моделі даних

Серед найсуттєвіших проблем реляційних баз даних, які на початку 1990-х років дали поштовх розвиткові альтернативних підходів до організації інформаційних ресурсів називають, зокрема, такі: по-перше, звуження методів і способів опрацювання даних до поняття таблиці, яке не є, загалом, еквівалентним до початкового поняття відношення і не завжди адекватно відображає логіку та семантику даних, по-друге, недостатньо коректне, з погляду змісту і застосування, подання невизначеностей за допомогою тризначної логіки та псевдоконстанти Null [1].

Незважаючи на активний розвиток таких новітніх підходів до організації та опрацювання інформаційних ресурсів як об'єктно-орієнтовані бази даних, слабкоструктуровані і напівструктуровані дані, web-ресурси, графічні та мультимедійні зображення тощо, реляційні бази даних продовжують залишатися основним засобом зберігання і опрацювання даних в інформаційних системах і технологіях різноманітного спрямування. Основними чинниками незмінної популярності реляційної моделі можна назвати такі:

- ґрунтовні теоретичні положення та прогресивні інформаційні технології роботи з реляційними базами даних забезпечують високу ефективність їх опрацювання;
- на сьогодні не існує методів і засобів опрацювання нереляційних даних які забезпечують однакову ефективність роботи з різними інформаційними ресурсами, сумірну з ефективністю застосування баз даних;
- функціональні можливості реляційної моделі не є вичерпаними, зокрема, вони не обмежені опрацюванням табличних структур їх може бути поширено на такі сфери як опрацювання слабкоструктурованих і неоднорідних даних, застосування об'єктно-орієнтованих та інших новітніх технологій.

Принципові положення, щодо додаткових можливостей реляційної моделі було викладено і обґрунтовано у Третьому маніфесті К. Дейта та Х. Дарвена [1]. Основною тезою цього документу є твердження, що реляційна модель у класичному трактуванні (не у версії SQL) має достатньо функціональних можливостей для вирішення проблем роботи з різнорідними, зокрема, нереляційними слабкоструктурованими даними, та застосування об'єктно-орієнтованих принципів у межах реляційної структури. Модель даних, яку пропонують у третьому маніфесті автори називають "істино реляційною моделлю". Особливістю реляційної моделі даних у поданні К. Дейта та Х. Дарвена є те, що її може бути застосовано для спільного подання, зберігання та опрацювання як реляційних так і нереляційних даних.

Базовими положеннями реляційної моделі, викладеними в [1] є такі.

*Значення і змінні.* Різницю між поняттями "значення" і "змінна" автори вважають принциповою і фундаментальної у процесах зображення та застосування даних.

*Скалярний тип даних.* Поняття скалярного (або, точніше, інкапсульованого) типу подається як певне узагальнення домену, і передбачає зображення одиничних елементарних значень даних у такий спосіб, який не потребує втручання користувача у їх внутрішню структуру при сприйнятті та застосуванні даних. Згідно з таким поданням, значенням скалярного типу можуть бути як число чи символьний рядок, так і бінарний файл, текст, XML-документ, web-сторінка або будь яка одиниця, над якою визначено дії, що виконують без втручання у її внутрішню структуру. Атрибут визначають як поіменовану визначену множину значень одного скалярного типу.

*Генерований тип кортеж.* Таки тип [1] застосовую як засіб утворення основної одиниці даних реляційної, призначенням якої є подання певних фактів. Розрізняють змінні типу кортеж та значення типу кортеж. Значенням типу кортеж є послідовність триплетів виду <A, T, v>, де

A – ім'я атрибута,

T – певний скалярний інкапсульований тип,

v – константа відповідного типу.

Множина впорядкованих пар виду <A, T> утворює схему (опис складу та структури) кортежу.

*Генерований тип відношення.* Автори [1] вводять поняття "змінної типу відношення" та "значення типу відношення". Значенням цього типу є множина значень типу "кортеж", які мають однакову схему. Схема кожного кортежу при збігається зі схемою відношення. Значення типу відношення застосовують як характеристику стану певної множини однотипних фактів, визначених у певній предметній області.

## 3. Фактологічна реляційна структура даних

Поняття факту в реляційній моделі не є новим. Вперше факт, як одиницю даних у відношеннях реляційної бази даних, було визначено у [2]. Згідно [2], поняття кортежу є занадто формалізованим і недостатньо відповідає суті та змісту даних. Це, як наслідок, створює низку проблем при роботі з таблицями (відношеннями), зокрема, в операціях вибору та оновлення даних. Саме тому у [2] запропоновано замінити кортеж поняттям факту як логічно завершеної достовірної змістовної одиниці, яка має власну інтерпретацію у визначеній предметній області. Кортеж (чи підкортеж)

відношення вважають константою, яка є зображенням деякого факту [2].

Згідно такого визначення один кортеж відношення може містити зображення множини фактів, кожен з яких має власну інтерпретацію. Така концепція цілком узгоджується з положеннями "істино реляційної моделі" у поданні К. Дейта та Х. Дарвена [1]. Наприклад, кортеж r відношення Студент з атрибутами № залікової, *Прізвище, Ім'я, Група, Середній_бал вигляду r=(12345, Петренко, Сергій, КН-41, 4.72)* є зображенням, зокрема, таких фактів:

- *Петренко Сергій* є студентом: $f_1$=*(Петренко, Сергій)* ∈ *Студент*,
- студент *Петренко* навчається в групі *КН-41*: $f_2$=(*Петренко, КН-41*) ∈ *Студент*,
- студент *Петренко* має середній бал 4.72: $f_3$=(*Петренко, 4.72*) ∈ *Студент*,
- студент *Петренко* має залікову книжку з номером *12345*: $f_4$=(*1234, Петренко*) ∈ *Студент* тощо.

Порівняння концепцій організації даних, викладених у [1] та [2] дозволяє поєднати їх у принципово новій моделі зображення та опрацювання даних. В основу підходу, що пропонується автором, покладено поняття факту, як множини значень, що подає певні достовірні відомості, релевантні щодо деякої предметній області.

Множину атрибутів, значення яких застосовують для зображення деякого факту *f* будемо називати його схемою і позначати як *sch(f)={*$A_{i1}$, $A_{i2}$, ..., $A_{ik}$*}*. Схема факту визначає його зміст та склад значень, з допомогою яких цей факт задано. Реалізацією факту *f* є деякий кортеж

$$r^f=<a_{i1}, a_{i2}, ..., a_{ik}>,$$

складений з припустимих значень атрибутів, що входять до схеми цього факту.

На основі такого визначення факту пропонується визначити фактологічне відношення (або відношення фактів) як структурну та функціональну одиницю даних. Попередньо визначимо поняття схеми фактологічного відношення. Схемою фактологічного відношення RF будемо називати вираз вигляду

$$Sch(R^F)=R^F(A_1, A_2, ..., A_n),$$

де $\{A_1, A_2, ..., A_n\}$ – множина атрибутів, які застосовують для подання певної категорії фактів. На відміну від схеми відношення у класичній реляційній моделі, схема фактологічного відношення є не формальним переліком властивостей класу сутностей, а системою семантичних координат (вимірів) деякого змістового простору. Кожен атрибут, в свою чергу, задає позначення виміру – ім'я, та його метрику – множину значень. У такому просторі можна визначати логічно завершені змістовні одиниці даних (факти) через значення вимірів. При цьому кількість значень, які застосовують для опису факту може бути різною.

Фактологічним відношенням $R^F$ зі схемою $Sch(R^F)$ будемо називати множину фактів, схема яких входить до складу схеми відношення

$$R^F=\{f \mid sch(f) \subseteq Sch(R^F)\}$$

У такий спосіб класичне поняття відношення реляційної бази даних значно розширюється і виходить за межі простого табличного зображення даних. Користувач отримує можливість за допомогою такого засобу в межах реляційної моделі баз даних оперувати структурами довільного вигляду, як показано, наприклад, на рис. 1.

*Відношення* $R^F$

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ | $A_7$ | $A_8$ |
|---|---|---|---|---|---|---|---|---|
| факт $f_1$ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| факт $f_2$ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| факт $f_3$ | | | | ■ | ■ | ■ | ■ | ■ |
| факт $f_4$ | | | ■ | ■ | ■ | ■ | | |
| факт $f_5$ | | | | | | | ■ | ■ |
| факт $f_6$ | | ■ | ■ | ■ | ■ | ■ | ■ | |
| факт $f_7$ | | ■ | ■ | ■ | ■ | ■ | ■ | |
| факт $f_8$ | | | | | | | | ■ |
| факт $f_9$ | | ■ | ■ | ■ | ■ | ■ | ■ | ■ |

Рис. 1. Приклад фактологічного відношення з кортежами різної розмірності

Запропонована фактологічна реляційна структура даних не заперечує класичну реляційну модель, оскільки стосується способів подання та опрацювання даних на зовнішньому рівні користувача. При цьому на концептуальному рівні бази даних зберігаються реляційні принципи роботи з даними, тобто утворюється комбінація "реляційна база даних – фактологічні засоби зовнішнього подання даних".

Джерелом значень для формування фактологічного відношення, у даному випадку, є таблиці (відношення) бази даних. Об'єкт, кортежі якого застосовують для утворення фактів, будемо називати *базовим відношенням*. Загалом, базове відношення може бути як таблицею, так і результатом перетворень однієї або більше таблиць бази даних. Загальну схему формування фактологічного відношення показано на рис. 2.

Як видно з рисунку, зовнішнє фактологічне відношення можна розглядати як результат виконання послідовності операцій проекції та селекції базового відношення.

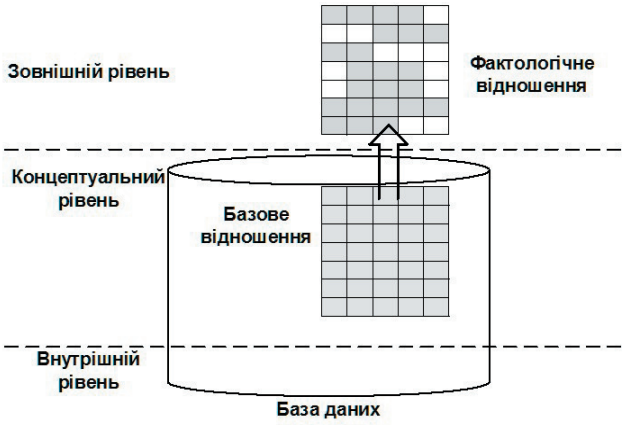


Рис. 2. Загальна схема формування фактологічного відношення на основі реляційної бази даних

Джерелом даних фактологічного відношення може бути не лише таблиця (відношення) реляційної бази даних. В загальному випадку, факт як змістовно завершену одиницю можна утворювати на основі значень,

поданих у інших форматах – тексту, XML-документів, web-ресурсів тощо. При цьому слід дотримуватися тих самих принципів, що і у випадку баз даних – кожен факт задається послідовністю значень, які можна однозначно інтерпретувати у визначеній предметній області.

На відміну від процедури формування фактів на основі кортежів базового відношення, утворення фактів з даних слабкоструктурованих форматів є значно складнішим. Ланцюжок значень, які зображають факт може бути сформовано способом, який враховує семантику поєднання цих значень у цілісне поняття. Таку процедуру, яка виділяє з джерела даних множину взаємопов'язаних значень та поєднує їх у факти будемо називати видобуванням фактів (fact mining). Методи і засоби вирішення цієї проблеми є окремим предметом досліджень і виходять за межі даної роботи.

Порядок та методи видобування фактів значною мірою залежать від виду та формату джерела даних. Загалом цей процес можна розглядати як частковий випадок таких методик як text mining, web mining, content mining тощо, які сьогодні є достатньо відомими і активно розвиваються. Загальну схему утворення фактологічного відношення на основі джерел довільних слабко структурованих форматів показано на рис. 3.

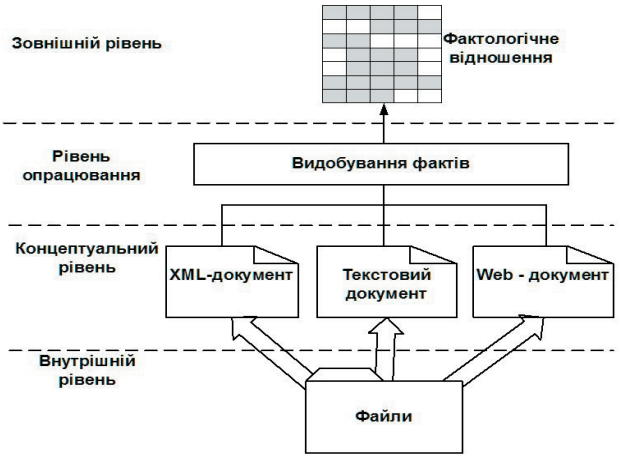


Рис. 3. Загальна схема формування фактологічного відношення на основі даних довільного формату

Можливості формування фактологічного відношення на основі структурованих та слабко структурованих джерел дозволяють використати їх як засіб інтеграції різнорідних даних на зовнішньому рівні. Результатом такого процесу є деяке фактологічне відношення $R^f$

$R^f=R^f_s$ UNION$^f$ $R^f_{ss}$,

де Rfs – фактологічне відношення утворене на основі базового відношення реляційної бази даних,

$R^f_{ss}$ – фактологічне відношення утворене на основі слабко структурованих джерел даних,

UNION$^f$ – оператор об'єднання фактологічних відношень, який об'єднує схеми та інформаційне наповнення двох фактологічних відношень. При цьому схема результуючого відношення утворюється об єднанням схем відношень операндів, а наповнення – об'єднанням наповнення цих відношень як показано на рис. 4.

У такий спосіб створюється можливість оперувати слабкоструктурованими неоднорідними даними, зберігаючи їх у як у структурованій реляційній формі так і у слабкоструктурованих форматах та даними з високим ступенем невизначеності. Окрім того, фактологічна реляційна структура може бути легко перетворена до інших форматів (наприклад XML, текстового тощо), виконуючи при цьому функції проміжної ланки між реляційними та слабкоструктурованими даними.

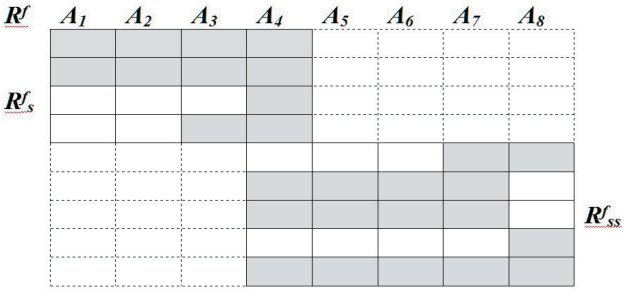


Рис. 4. Схема виконання операції об'єднання фактологічних відношень

## Висновок

Застосування фактологічних відношень як засобу зовнішнього зображення різнорідних даних, початково поданих і збережених у реляційних структурах, дозволяє отримати такі переваги:

- перейти від структурного принципу подання та опрацювання даних до семантичного;
- застосовувати для зображення одиниць даних кортежі різної розмірності;
- відмовитись від застосування тризначної логіки та псевдо-константи Null ;
- поєднати у єдиному середовищі структуровані та напісвтруктуровані форми подання даних.

Підхід, запропонований автором, може бути застосовано у розв'язанні, зокрема, таких проблем як інтеграція інформаційних ресурсів, інтелектуальних аналіз даних, опрацювання неповних і неточних даних, перетворення форматів, створення гетерогенних структур даних тощо.

Література

1. Date C.J. Foundation for Future Database Systems: The Third Manifesto, 2nd edn./ Date C.J., Darwen H.- Harlow: Addison Wesley Longman, 2000.
2. Desai B.C. Fact structures and its application to updates in Relational Databases./ Desai B.C., Goyal P., Sadri F. // Information Systems Vol. 12, No 2, 1987.- p. 215-221.
3. Калиниченко Л.А. Методы и средства интеграции неоднородных баз данных. / Леонид Калиниченко. – М.- Наука, 1983.- 424 с.