МЕТОД СЕМАНТИЧНОЇ ІНТЕГРАЦІЇ ЛОКАЛЬНО НЕЗАЛЕЖНИХ ДАНИХ

Руденко Д.О.
Харківський національний університет радіоелектроніки

Однією з проблем у розвитку інформаційних систем з розподіленою обробкою даних є семантична інтеграція баз даних (БД), зокрема, визначення еквівалентності атрибутів інтегрованих БД. Порівняння всіх можливих пар атрибутів є невиправдано трудомістким завданням, тим більше що більшість пар не містять однакову інформацію. Проте, можна припустити, що деякі атрибути будуть не враховані при відображенні на одному глобальному атрибуті системи, що інтегрується.

Семантика БД може бути втілена в модель даних, концептуальну схему, прикладні програми тощо Семантичні відмінності між атрибутами можна вважати невідповідністю або неоднорідністю в залежності від програми, тому часто важко визначити і усунути всю семантичну неоднорідність.

Рішення подібних проблем в разі неоднорідності (як структурної, так і семантичної) корпоративних інформаційних систем неможливо без побудови адекватного інтегрованого уявлення системних метаданих на основі моделей даних, інваріантних до структурних і архітектурних особливостей [1]. При цьому моделі даних повинні носити розширюваний, динамічний характер і допускати необмежене розширення за допомогою метауровневих переходів. Моделі повинні включати математичні засоби, як для опису даних, так і для маніпулювання даними. Кожна модель повинна бути забезпечена відповідними інструментальними засобами автоматизованого проектування, що дозволяють забезпечити конвеєрну інтеграцію семантично цілісних додатків.

Для розробки семантичної процедури інтеграції на основі специфікацій метаданих та концептуальної схеми БД необхідно мати можливість визначати приналежність атрибутів до деякого класу інтегрованої даних, а також адаптувати знання, отримані в процесі семантичного узгодження неоднорідності. Визначення еквівалентності атрибута передбачає інтеграцію, в першу чергу, на основі їх доменних відносин: рівні (equal), містить (contains), що перекривають (overlap), що містяться в (contained-in) і не перетинаються (disjoint). Визначення таких відносин є досить трудомістким. Так, наприклад, якщо схема БД має 100 типів сутностей і в середньому п'ять атрибутів для кожного типу сутностей, то необхідно проаналізувати 250000 пар атрибутів (для кожного атрибута в одній схемі, повинні бути проаналізовані всі атрибути в інших схемах). Крім цього, як правило, виникає ще одна суттєва проблема - це погана стійкість до помилок. Невелика кількість невірних даних може привести інтегровану систему до неправильних висновків про узгодженість доменів.

На практиці, для ідентифікації пар типів атрибутів і типів відносин при побудові інтегрованої схеми використовуються евристичні методи, пов'язані доменними відносинами: equal, contains, overlap і contained-in. З іншого боку, таке завдання може бути вирішене без використання евристики (тобто,

автоматизовано) для пар атрибутів, пов'язаних розглянутими доменними відносинами, крім відносини «перетинаються» (intersect).

Специфікації атрибутів на рівні схеми являють собою основні типи даних, довжину і додаткові типи даних, такі як формат і обмеження (первинні ключі, зовнішні ключі, діапазон допустимих значень, заборона порожніх значень, а також обмеження доступу). Багато реляційних БД зберігають цю інформацію в таблицях метаданих, дозволяючи запитам SQL витягати необхідну інформацію. Таким чином, можна розробити аналізатор для отримання цієї інформації.

Для неоднорідних систем дані атрибутів, як правило, відрізняються, навіть якщо їх властивості, такі як тип даних і розмір, збігаються. В основному це пов'язано з різними моделями визначення даних.

Очевидно, що не існує ідеальної процедури або стандартного набору правил, які вирішать проблему визначення семантичної еквівалентності атрибутів в неоднорідних БД, так як відносини між атрибутами неоднозначні, а доступ до інформації БД змінюється в процесі функціонування інформаційної системи. Для вирішення завдання визначення семантичної близькості різних атрибутів можна скористатися методом розпізнавання шаблонів, заснованим на нейронних мережах. Даний метод дозволяє визначити ступінь подібності атрибутів безпосередньо за їх значенням і емпірично отримати результат без попереднього аналізу закономірностей в наборі даних. Переваги використання нейронних мереж для визначення еквівалентності атрибутів в порівнянні з методами, заснованими на використанні фіксованих правил, полягають у наступному:

- нейронні мережі вирішують такі завдання, як класифікація і узагальнення без визначення відповідних правил, так як нейронні мережі можуть навчатися і, як наслідок, адаптуються до змін в примірниках БД;
- початкові значення можна змінювати динамічно відповідно до вхідних даних;
- нейронні мережі можна узагальнити через їх можливості правильно реагувати на дані, невикористовувані в процесі навчання.

Узагальнення грає важливу роль в розглянутих задачах, так як вхідні дані: імена, специфікації і значення атрибутів в неоднорідних БД часто бувають зашумлені і\або неповні. Побудова нейронної мережі будується на основі структурних та інформаційних характеристик БД. По-перше, наявна інформація з різних БД використовується в якості вхідних даних для алгоритму таблиці класифікації атрибутів, що самоорганізується. По-друге, вивід класифікатора використовується в якості навчальних даних для аналізу категорій і алгоритму розпізнавання, після чого навчений алгоритм визначає подібність між парами атрибутів в різних БД.

1. Maglott D., Ostell J., Pruitt K.D., Tatusova T. Entrez Gene: gene-centered information at NCBI // Nucleic Acids Research, 2005. Vol. 33 (Database Issue). P. D54–D58. http://www.pubmedcentral.nih.gov/articlerender.fcgi? artid=539985].