УДК 004.65:004.421

АВТОМАТИЗОВАНЕ ВИЯВЛЕННЯ НАДЛИШКОВИХ І ДУБЛЬОВАНИХ ДАНИХ У РЕЛЯЦІЙНИХ БАЗАХ ДАНИХ

Колін Д. П.[1]; Озарінська О. В.[2]
[1,2]Національний університет «Одеська політехніка»
[1,2]Україна, Одеса
[1]dan.kolin@gmail.com; [2]o.v.ozarinska@op.edu.ua

Анотація. *Проаналізовано методи виявлення точних і наближених дублікатів у реляційних базах даних. Розглянуто вплив нормалізації значень, вибору атрибутів і попереднього відбору пар на якість зіставлення. Обґрунтовано критерії порівняння підходів за точністю, повнотою та обчислювальними витратами. Визначено обмеження автоматичного об'єднання записів і напрями експериментального дослідження.*

Ключові слова: *реляційні бази даних; дублікати; якість даних; зіставлення записів; очищення даних.*

Вступна частина. У сучасних інформаційних системах реляційні бази даних накопичують відомості з різних джерел: вебформ, зовнішніх сервісів, імпортованих таблиць і внутрішніх модулів. Повторне внесення інформації, відмінності у форматуванні та помилки введення призводять до появи дубльованих записів. Наслідками стають спотворення аналітичних показників, зайві витрати на зберігання, некоректні результати пошуку та ускладнення підтримки цілісності даних.

Необхідно розрізняти точні дублікати, коли значення суттєвих атрибутів збігаються, і семантичні дублікати, коли записи описують той самий об'єкт, але містять відмінності в написанні, скороченнях або окремих реквізитах. Надлишковість є ширшим поняттям: вона може виникати внаслідок невдалого проєктування схеми, повторного зберігання залежних відомостей або дублювання сутностей. Для боротьби з різними видами надлишковості потрібні різні засоби – від нормалізації структури до зіставлення значень [1].

Мета роботи – провести порівняльний аналіз підходів до виявлення дубльованих записів у реляційних базах даних і визначити умови, за яких точне або наближене зіставлення забезпечує прийнятний баланс між точністю, повнотою та витратами обробки. Дослідження має аналітичний характер; кількісні висновки потребують окремої експериментальної перевірки.

Основна частина. Об'єктом аналізу є записи реляційних таблиць, що містять ідентифікаційні та описові атрибути: назви організацій, прізвища, адреси, номери телефонів, електронні адреси та інші реквізити. Наявність однакових значень в одному полі не завжди означає дублювання сутності. Наприклад, декілька клієнтів можуть мати спільну адресу, а одна особа – кілька контактних номерів. Тому правило зіставлення повинно враховувати семантику даних і сукупність ознак.

У задачі виявлення дублікатів принципово відрізняються два підходи. Перший ґрунтується на точній рівності нормалізованих значень і придатний для записів зі стабільними ідентифікаторами. Другий використовує міри подібності, що дозволяють виявляти записи з друкарськими помилками, скороченнями та різним порядком слів. Жоден із підходів не є універсальним: жорсткі правила знижують ризик помилкового об'єднання, але пропускають частину повторів, тоді як гнучкі правила розширюють охоплення ціною додаткових хибних збігів.

Значну роль відіграє семантика атрибутів. Первинний ключ забезпечує унікальність рядка, але не гарантує унікальності описуваної сутності. Натомість реєстраційний номер або інший надійний зовнішній ідентифікатор може бути вагомішою ознакою відповідності, ніж

текстова назва. Контактні дані, адреса та найменування мають різну доказову силу залежно від предметної області. Отже, однакові правила зіставлення не слід механічно переносити між базами клієнтів, підприємств і товарів.

Нормалізація текстових даних суттєво впливає на результати обох підходів. Усунення зайвих пробілів, уніфікація регістру та форматів телефонів дають змогу знаходити повтори без складних алгоритмів. Проте надмірне перетворення може знищити значущі відмінності: вилучення частин назви, цифр або скорочень іноді призводить до збігу різних об’єктів. Особливої уваги потребують українські назви, апостроф, дефіси, варіанти транслітерації та абревіатури. Тому вихідні значення необхідно зберігати, а правила перетворення – документувати.

Інше важливе питання – обсяг порівнянь. Повне зіставлення кожного запису з усіма іншими швидко стає ресурсомістким зі збільшенням таблиці. Попереднє групування за окремими атрибутами дозволяє скоротити кількість кандидатних пар, однак створює ризик пропуску дублікатів, що потрапили до різних груп. Наприклад, групування за першою літерою назви буде неефективним за наявності помилок саме на початку слова. У спеціалізованих дослідженнях пошуку відповідностей між записами цей компроміс розглядається як одна з центральних проблем масштабування [2].

Точне зіставлення має переваги простоти, прозорості та відносно невеликих обчислювальних витрат. Його доцільно застосовувати до уніфікованих реквізитів, для яких припустима сувора перевірка рівності. Наближене зіставлення використовує, зокрема, редакційну відстань, порівняння наборів слів або символьних фрагментів. Воно корисне для текстових атрибутів, що часто вводяться вручну, проте потребує вибору порогу подібності. Занадто низький поріг збільшує кількість хибних збігів, а занадто високий – кількість пропущених повторів [3].

Перспективним є комбінування різнорідних ознак. Схожість найменувань може підсилюватися збігом адреси, електронної пошти або іншого реквізиту. Водночас суперечність у надійному ідентифікаторі повинна зменшувати впевненість у відповідності навіть за високої текстової подібності. Такий підхід дозволяє перейти від правила «однакові рядки» до оцінки того, чи описують записи одну сутність. Його обмеженням залишається залежність від якості та заповненості атрибутів.

Для коректного порівняння методів необхідно розділяти точність і повноту. Точність характеризує частку справжніх дублікатів серед виявлених пар, а повнота – частку знайдених дублікатів серед усіх наявних у контрольному наборі. Висока повнота не компенсує низької точності в системах, де помилкове об’єднання може спричинити втрату важливої інформації. Доцільно також враховувати кількість кандидатних пар, тривалість обробки, використання оперативної пам’яті та складність пояснення рішення користувачеві.

Результат виявлення збігів не тотожний рішенню про видалення записів. Автоматичне об’єднання допустиме лише за достатньої надійності ідентифікації та після перевірки зв’язків між таблицями. У неоднозначних випадках доцільно формувати рекомендації для ручного підтвердження. Такий розподіл відповідальності важливий для інформаційних систем, де один запис може бути пов’язаний із замовленнями, платежами або історією змін. Відтак ефективність пошуку дублікатів слід оцінювати разом із ризиком подальшого некоректного очищення.

Для ілюстрації запропонованого підходу розглянемо умовні записи «ТОВ Альфа-Сервіс» і «Альфа Сервіс, ТОВ». Після нормалізації організаційно-правової форми та розділових знаків їхні назви можуть виявитися близькими, однак остаточний висновок потребує порівняння адреси, контактних даних або іншого надійного реквізиту. Натомість записи з однаковою назвою, але різними реєстраційними ідентифікаторами не можна автоматично вважати дублікатами. Цей приклад демонструє логіку перевірки, а не результати проведеного експерименту.

Порівняння підходів за обчислювальними витратами також не зводиться до часу одного зіставлення. У великих таблицях значну частину роботи становлять пошук

кандидатних пар, завантаження атрибутів та зберігання проміжних результатів. Точне зіставлення за індексованими полями зазвичай простіше масштабувати, тоді як наближені текстові порівняння можуть вимагати додаткової фільтрації. Отже, вибір методу залежить від обсягу даних, частки помилкових записів, доступної пам'яті та допустимого часу виконання.

Для перевірки сформульованих положень доцільно створити контрольний набір синтетичних записів із відомими відповідностями та типовими спотвореннями: перестановкою слів, пропусками символів, різним регістром і варіантами форматування. Предметом порівняння можуть стати точне зіставлення після нормалізації, наближене текстове зіставлення та комбінований підхід із попереднім відбором кандидатних пар. За однакових вихідних даних слід вимірювати точність, повноту, кількість порівнянь і час обробки. Така постановка експерименту дозволить визначити не лише найшвидший, а й найбільш збалансований підхід.

Проведений аналіз показує, що якість виявлення дублікатів визначається не окремим алгоритмом, а відповідністю способу зіставлення властивостям даних. Для структурованих і надійних ідентифікаторів перевагу має точна перевірка; для текстових полів із типовими помилками потрібні наближені міри подібності; для великих наборів даних критичним стає обмеження кількості кандидатних пар. Таким чином, основним предметом оптимізації є баланс між достовірністю встановлення відповідності та витратами на її перевірку.

Висновки. Порівняно точне, наближене та комбіноване зіставлення записів у реляційних базах даних. Встановлено ключові чинники вибору підходу: надійність атрибутів, характер помилок, обсяг таблиць і допустимий ризик хибного об'єднання. Обґрунтовано необхідність спільного оцінювання точності, повноти та обчислювальних витрат, оскільки покращення одного показника не гарантує загального підвищення якості очищення даних.

Практична цінність аналізу полягає у визначенні умов застосування різних способів виявлення дублікатів без прив'язки до конкретної системи керування базами даних. Подальше дослідження передбачає експериментальне порівняння методів на контрольованих наборах записів.

Використання інструментів ШІ: не використовувались.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4), 3–13.

[2] Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin, Germany: Springer.

[3] Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 1–16. https://doi.org/10.1109/TKDE.2007.250581

Kolin D. P., Ozarinska O. V.

Automated detection of redundant and duplicate data in relational databases

Abstract. *The paper compares exact, approximate, and combined approaches to duplicate record detection in relational databases. Particular attention is paid to attribute normalization, candidate pair selection, and the trade-off between precision, recall, and computational cost. The conditions under which different matching methods are appropriate are discussed, together with the risks of incorrect record merging. A controlled experimental comparison is outlined as future work.*

Keywords: *relational databases; duplicate detection; data quality; record linkage; data cleaning.*