УДК 004.912

ВИЯВЛЕННЯ БРУДНИХ ДАНИХ У СХОВИЩАХ ДАНИХ

Дідковська М.В., Гранаткіна Т.Д.
Навчально-науковий комплекс
«Інститут прикладного системного аналізу»,
Національний технічний університет України
«Київський Політехнічний Інститут»

Досліджена проблема виявлення брудних даних у сховищах даних. Проведено порівняльний аналіз функцій мір подібності записів. Розглянуто підходи до виявлення брудних даних. Запропоновано алгоритм методу дедублікації даних.
Ключові слова: сховища даних, брудні дані, дедублікація, очищення даних, міра подібності, якість даних.

Сьогодні системи сховищ даних є ключовими факторами технологічної інфраструктури корпоративної інформації. За рахунок консолідації даних з різних джерел в центральне сховище даних, корпорації мають можливість використовувати додатки для аналізу даних і отримання інформації, що має стратегічне та тактичне значення для їхнього бізнесу. Ці додатки засновані на використанні бізнес-аналітики, отриманої зі сховищ даних або баз даних, і надають велике значення високій якості даних. Аналітики витрачають значні ресурси на пошук, виправлення або будь-який інший спосіб ліквідації проблем з даними. Як правило, більше 80% часу, присвяченого аналітичним проектам, витрачається на обробку і очищення брудних даних [1].

Таким чином, якість даних має бути однією з ключових ініціатив кожної компанії, яка використовує великий масив даних для прийняття важливих бізнес-рішень. В області сховищ даних, очищення даних особливо необхідне при об'єднанні декількох баз даних. Записи, що відносяться до однієї і тієї ж сутності можуть бути представлені в різних форматах і різними наборами даних, або представлені помилково. Таким чином, постає питання про виявлення та усунення таких дубльованих записів. Ця задача відома як merge/purge problem.

Головним проявом брудних даних є існування дублікатів, тобто декількох записів у базі даних, – яка може бути як автономною, так і інтегрованою, – що відносяться до одного і того ж об'єкту реального світу.

Якщо декілька записів відносяться до однієї і тієї ж сутності реального світу, але синтаксично записи не є однаковими, то можемо вчиняти двома способами. Можна розглядати один із записів як правильний, а інший записів як дублюючий, що містить помилкову інформацію. Тоді мета – очищення бази даних від повторюваних записів. Альтернативою є розгляд кожного окремого запису в якості часткового джерела інформації. Тоді мета полягає в об'єднанні дублюючих записів для отримання одного запису з більш повною інформацією [2].

Незважаючи на ряд існуючих розроблених підходів до вирішення даної проблеми, вони мають певні недоліки: у роботі [1] описаний алгоритм очищення бази даних використовує зовнішні файли-джерела для валідації правильності даних; в алгоритмі роботи [3] виділяються такі його обмеження, як застосування алгоритму тільки до полів з назвами та значний об'єм ручної роботи на стадії препроцессингу; інші методи [4] обмежуються прямолінійним застосуванням міри подібності до одного поля і не аналізують запис в цілому.

Головною метою цієї роботи є проведення аналізу мір подібності та підходів до виявлення брудних даних з подальшою розробкою алгоритму методу виявлення дублюючих записів, призначеного для застосування у сховищах даних.

Сформулюємо задачу виявлення подібності двох рядків. Якщо дано два набори рядків X та Y, необхідно знайти всі пари рядків (x, y), $x \in X$, $y \in Y$ такі, що x та y відносяться до одного й того ж об'єкту реального світу. Такі пари називаються відповідностями. Для вирішення проблеми точності вводиться поняття міри подібності.

Нехай $s=s(x,y)$ – функція міри подібності така, що для кожної пари (x, y) повертає оцінку в інтервалі від 0 до 1. Тоді вважаємо, що між x та y існує відповідність, якщо $s(x,y) \geq t$, де t – вибране порогове значення. Чим вища оцінка, тим більша ймовірність відповідності даних рядків.

$s:(x,y) \rightarrow l, l \in [0,1]$

Виділяють чотири групи міри подібності: послідовні (розглядають рядок як послідовність символів), групові (розглядають рядок як набори символів), гібридні та фонетичні (в основі яких лежить фонетичне звучання слів).

До найвідоміших послідовних функцій подібності відноситься міра Левенштейна та її узагальнення: міри Нідлмана-Вунша і Affine gap. Остання дозволяє ефективно обробляти довгі пропуски [5].

Міра Левештейна визначає найменшу вартість $d(x,y)$ трансформування рядка x в рядок y. Перетворення рядка відбувається застосуванням до літер послідовності операторів: вилучення, вставка, заміна. Міра подібності має вигляд: $s(x,y)=d(x,y) / max(length(x),length(y))$.

Міра Жаккара є найбільш використовуваною груповою мірою. Нехай B_x та B_y – набори n-грам, згенерованих з рядків x та y, тоді міра Жаккара визначається як: $J(x,y)=|B_x \cap B_y| / |B_x \cup B_y|$.

Міра TF-IDF використовує поняття оцінки, що застосовується в інформаційному пошуку для знаходження релевантних документів за ключовими словами. Основна ідея полягає у тому, що два ряді являються подібними, якщо містять однакові характерні терміни. Кожен рядок $x_i=a_i \ldots a_k$ перетворюється на набір термінів $B_{xi}=\{a_i, \ldots, a_k\}$. Для кожного терміну t та документу d обчислюється $tf(t,d)$ як відношення числа входжень терміну t до загальної кількості термінів документу d. Для кожного терміну t обчислюється $idf(t)$ як відношення загальної кількості документів в колекції до кількості документів, що містять термін t. Далі кожному документу d співставляється веткор v_d: $v_d(t)=tf(t,d) \cdot idf(t)$. Нехай x та y – порівнювані рядки, T – набір всіх термінів в колекції. Тоді оцінка подібності рядків визначається як:

$$s(x,y) = \frac{\sum_{t \in T} v_x(t) \cdot v_y(t)}{\sqrt{\sum_{t \in T} v_x(t)^2} \cdot \sqrt{\sum_{t \in T} v_y(t)^2}}$$

Гібридні міри, до яких відносяться Soft TF-IDF та Узагальнена міра Жаккара, мають в своїй основі

оригінальні алгоритми (TF-IDF та Жаккара відповідно), але модифіковані таким чином, щоб знаходити не точні, але схожі терміни або н-грами. Вони мають більш високу складність, але дають порівняно кращі результати [5].

Співставимо найуживаніші міри подібності між собою для порівняння в таблиці 1.

Виділяють три основні групи методів, що можуть виявляти подібні записи, застосовуючи оцінку подібності:

- методи на основі правил;
- методи на основі навчання з вчителем;
- методи на основі навчання без вчителя.

Кожен з них має свої переваги і недоліки, що відіграють важливу роль у виборі того чи іншого методу для конкретної поставленої задачі.

Ідея методів на основі правил полягає у тому, щоб розглядати характеристики об'єктів та їх вплив на об'єкт в цілому згідно з встановленими користувачем правилами. В даному випадку обчислюється загальний показник подібності, наприклад, у вигляді лінійної комбінації функцій подібності характеристик об'єкту з ваговими коефіцієнтами або у вигляді правила логістичної регресії. Вагові коефіцієнти тут показують ступінь впливу схожості даної характеристики об'єкта в порівнюваних записах на подібність його в цілому.

Методи, в основі яких лежить навчання з вчителем вимагає мати досить великий набір тренувальних даних, що складаються з набору характеристик об'єкта двох записів та позначку, яка індикує чи є записи подібними. Таким чином, модель навчається

Таблиця 1

Порівняльний аналіз мір подібності

Міра	Тип міри	Призначення	Переваги	Недоліки	Складність
Левенштейн	Послідовна	Друкарські помилки в результаті помилкового введення	дає ефективні результати в разі допуску не більше 2 друкарських помилок	не підходить для виявлення інших видів помилок, наприклад, для випадків написання різних форм імен або пропущених частин назв об'єктів; порівняно низька швидкість роботи	$O(n^2)$
Нідлман-Вунша	Послідовна	Друкарські помилки в результаті помилкового введення	дозволяє гнучке оброблення порівняння літер в словах.	Аналогічні до міри Левенштейна	$O(n^2)$
Affine Gap	Послідовна	Пропущені частини імен, прізвищ або назв об'єктів в результаті слабкої стандартизації полів	здатний обробляти довгі пропуски в словах, наприклад, пропущені частини імен через різні стандарти написання	дає погані результати на коротких словах та при довгих пропущених словах; порівняно великий час роботи	$O(n^2)$
Жаро-Уінклера	Послідовна	В першу чергу для коротких імен та прізвищ	відносно висока швидкість роботи алгоритму, показує гарні результати як при друкарських помилках, так і пропущених словах.	іноді вважає схожими різні слова, якщо вони містять певну кількість однакових літер у тому ж порядку	$O(n^2)$
Жаккара	Групова	Друкарські помилки в результаті помилкового введення, скорочення слів, варіації назв	ефективно працює при наявності коротких пропущених слів; порівняно висока швидкість роботи.	дає погані результати якщо значна частина слова відсутня або відсутні слова є довші за присутні.	$O(n^2)$
TF/IDF	Пошуку інформації	Виявлення розпізнавальних термінів у записах зі списку	гарні результати за присутності посад осіб, титулів та ін. загальних слів, що не повинні впливати на порівняння рядків	не підходить до застосування у випадках друкарських помилок, злитого написання слів	$O(\|n\|\|m\|)$
Soft TF/IDF	Пошуку інформації, нечітка	Виявлення розпізнавальних термінів у записах з можливим неправильним написанням	ті ж, що і для міри TF-IDF і здатність виявляти друкарські помилки.	висока складність алгоритму	$O(n^2\ m^2)$
Узагальнена Жаккара	Групова, нечітка	Виявлення текстуально подібних структур з можливим неправильним написанням	Аналогічні soft TF/IDF	висока складність алгоритму	$O(n^4)$
Soundex	Фонетичні	Виявлення фонетично подібних структур	висока швидкість роботи; можливість розпізнавати помилки, що були допущені при отриманні даних усним шляхом	Не працює для випадків друкарських помилок	$O(n^2)$

Джерело: розроблено авторами та [5]

розпізнавати подібність яких характеристик впливають на подібність об'єкту в цілому. На відміну від методів на основі правил, в даному випадку це робиться автоматично. Це є зручним у разі наявності великої кількості характеристик об'єкту, тобто, полів у таблиці. В якості навчального алгоритму можуть бути застосовані дерева рішень або метод опорних векторів [5].

До методів на основі навчання без вчителя відноситься кластеризація. Такі методи, як агломераційна ієрархічна кластеризація, k-means є широко використовувані для знаходження брудних даних. Порівнювані об'єкти, а пізніше групи об'єктів, розміщуються всередині кластерів таким чином, щоб подібність між об'єктами в середині кластерів була високою, а між кластерами – низькою. Зі способів підрахунку подібності кластерів виділимо: єдиний зв'язок, повний зв'язок, середній зв'язок та канонічний запис. Останній означає вибір запису, що буде представляти весь кластер.

Наведемо основні переваги та недоліки розглянутих методів у Таблиці 2.

Таблиця 2

Переваги та недоліки методів виявлення брудних даних

Метод	Переваги	Недоліки
На основі правил	1. Висока швидкість роботи 2. Високий рівень налаштованості користувацького інтерфейсу	1. Бінарне відношення записів
Кластеризація (навчання без вчителя)	1. Порівняно нижча швидкість роботи	1. Кожен кластер представляє окрему сутність
На основі навчання з вчителем	1. Вимагає мати великий тренувальний набір	1. Можливе автоматичне визначення вагових коефіцієнтів

Джерело: розроблено авторами за даними [5]

Таким чином, проаналізувавши переваги і недоліки підходів до знаходження брудних даних, сформулюємо наступний алгоритм методу їх виявлення для випадку сховищ даних:

1. Встановити з'єднання з базою даних.
2. Відсортувати цільову таблицю згідно з вибраним ключем.
3. Виконати команду select таблиці з бази даних.
4. Визначити ключові константні значення: розмір ковзного вікна, вагові коефіцієнти кожної характеристики таблиці, порогові значення для кожного методу (за замовченням порогові значення мають рекомендовані значення).
5. Підрахувати подібність записів для кожної вибраної колонки таблиці, використовуючи обраний метод, згідно з нашими рекомендаціями:
   - для особистих імен рекомендується використовувати метод на основі міри Жаккара, Жаро-Уінклера або Affine gap;
   - для імен компаній рекомендується використовувати метод на основі міри TF-IDF; у разі порівняно невеликого об'єму даних, Soft TF-IDF;
   - для числових полів, наприклад, номерів телефонів, рекомендується використовувати метод на основі міри Левенштейна;
   - для нестандартизованих категоріальних полів рекомендується використовувати метод на основі міри Нідлмана-Вунша, Жаккара або TF-IDF.
6. Підрахувати інтегровану матрицю подібності, використовуючи вагові коефіцієнти.
7. Створити таблицю результатів і виконати команду insert результуючих даних.
8. Від'єднатися від бази даних.

В роботі був вибраний метод на основі правил користувача з двох причин. По-перше, він показує високу швидкість роботи, що є вкрай важливим при великих обсягах даних. По-друге, він може мати високий рівень налаштованості користувацького інтерфейсу, тобто вибір порогових значень, вагових коефіцієнтів відбувається людиною. Оскільки цільова група користувачів є бізнес-аналітики, це надає гнучкості програмі і її параметри можуть бути підібрані в залежності від особливостей компанії, таблиці та даних, що в ній аналізуються.

Таким чином, в роботі були розглянуті різні типи мір подібності і виділені їх переваги та недоліки, що дозволило визначити які з них будуть придатні для застосування у методах дедублікації у сховищах даних. Основним критерієм була швидкість роботи алгоритму при великих об'ємах даних. Також, на відміну від існуючих алгоритмів, наш метод майже не потребує попередньої ручної обробки даних і може розглядати як окремі поля таблиці, так і записи в цілому.

У майбутніх дослідженнях можлива подальша оптимізація алгоритму з метою збільшення швидкості роботи програмного продукту. Одним з можливих варіантів є впровадження паралельних процесів виконання програми, застосування технології map-reduce.

Список літератури:

1. Francis L.A. Dancing with Dirty Data. Mathods of Exploring and Cleaning Data // Casualty Actuarial Society Forum. – 2005. – P. 198-248.
2. Lee M. L. Cleansing Data for Mining and Warehousing. – Singapore. – 10 p.
3. Porwal S.,Vora D. A Comparative Analysis of Data Cleaning Approaches to Dirty Data // International Journal of Computer Applications. – Vol. 62. – No. 17. – January 2013. – P. 30-34.
4. Cohen W.W., Ravikumar P., Fienberg S.E. A Comparison of String Distance Metrics for Name-Matching Tasks. – 6 p.
5. Doan A. et al. "String Matching" in Principles of Data Integration, ed. Waltham: Morgan Kaufmann. – 2012. – P. 95-110.

Дидковская М.В., Гранаткина Т.Д.
Учебно-научный комплекс «Институт прикладного системного анализа»,
Национальный технический университет Украины
«Киевский Политехнический Институт»

ВЫЯВЛЕНИЕ ГРЯЗНЫХ ДАННЫХ В ХРАНИЛИЩАХ ДАННЫХ

Аннотация
Исследована проблема выявления грязных данных в хранилищах данных. Проведен сравнительный анализ функций мер подобия записей. Рассмотрены подходы к выявлению грязных данных. Предложен алгоритм метода дедубликации данных.
Ключевые слова: храналише данных, грязные данные, дедубликация, очистка данных, мера подобия, качество данных.

Didkovska M.V., Granatkina T.D.
Educational and Scientific Complex «Institute for Applied Systems Analysis»
National University of Ukriane
«Kyiv Polytechnic Institute»

DIRTY DATA DETECTION IN DATA WAREHOUSES

Summary
Dirty data detection problem in data warehouses was studied. Comparative analysis of similarity measure functions was performed. Methods of dirty data detection were reviewed. Algorithm for deduplication method was proposed.
Keywords: data warehouse, dirty data, deduplication, data cleansing, similarity measure, data quality.

УДК 004.932.72

МЕТОД ЛОКАЛІЗАЦІЇ СТРУКТУРНИХ ОБ'ЄКТІВ НА ФОТОГРАФІЇ

Мельник В.В.
Національний технічний університет України
«Київський політехнічний інститут»

Запропоновано метод локалізації структурних об'єктів на фотографіях з використанням нелінійної моделі. Метод застосовано для знаходження ключових точок тіла людини. Для даної задачі встановлено покращення метрик якості.
Ключові слова: структурні об'єкти, локалізація, нелінійна модель.

Одними з основних задач комп'ютерного зору є виявлення та локалізація об'єктів. Ці задачі слугують базовим блоком для алгоритмів сегментації, відстеження об'єктів та розпізнавання дії на відео.

Рис. 1. Структурна модель
Джерело: розроблено автором

Структурні об'єкти являються найбільш складним випадком для локалізації. Під структурним об'єктом ми будемо розуміти множину ключових точок та апріорні закони їх розміщення в тривимірному просторі. Дані точки являються візуально інваріантними, а їх кількість – фіксованою. Внаслідок рухливості ключових точок, їх відносне положення на площині фотографії не є фіксованим[3]. Процес проектування на площину також створює додаткові деформації розмірів частин об'єкту. Крім того, частина точок може бути перекрита іншими елементами сцени або об'єкта. Найбільш важливі приклади структурних об'єктів: кисть, обличчя, тіло людини. Знаходження елементів кисті необхідно для розпізнавання жестів людини. Локалізація ключових точок обличчя застосовується для обчислення «хешу» – індивідуальної інваріантної до проектування та поворотів характеристики, яку зручно використовувати для задачі ідентифікації. Виявлення положення частин тіла людини важливе для створення сучасних людино-машинних інтерфейсів, розпізнавання жестів. В даній роботі метод локалізації структурних об'єктів буде застосовано для пошуку ключових точок людини.

Структурною моделлю об'єкту будемо називати множину ключових точок та фіксований ациклічний однозв'язний граф, побудований на цих точках [1]. Даний граф задає взаємозв'язки між точками (Рис. 1).