УДК 004.658.4

Аспірант Комісар Д.А., студент Семенякін В.С.

**Національний технічний університет України
«Київський політехнічний інститут»**

СПОСІБ ВИПРАВЛЕННЯ ЗАПИСІВ БАЗ ДАНИХ ТИПУ «ПРІЗВИЩЕ ІМЯ ПО-БАТЬКОВІ» З ВИКОРИСТАННЯМ ЗВАЖЕНИХ СЛОВНИКІВ

Abstract

Dmistriy A. Komisar,graduate student,Volodymyr Semenyakin, student

Method of correcting "Name / Second-Name / Patronymic" data base fields, using weighted dictionaries

This paper presents an automated method of correcting the information fields of the database for the subsequent unification of records. The possible structure of the field and possible distortion was examined. Particular attention was paid to the separation of connected words. A way of teaching the system to improve the accuracy of the algorithm is proposed. Benefits of using the method were presented.

Вступ

Задачею переважної більшості баз даних є опис певних взаємозв'язків, що стосуються інформації про людей. Подібний опис потребує збереження особистих даних людей, перш за все, їх прізвища, ім'я та ім'я по-батькові (далі – ПІП). Однією з найпоширеніших причин виникнення помилок в подібних БД є механізм їх створення – вручну, без верифікації. З іншого боку, все активніше застосовуються технології розпізнавання образів, котрі можуть спотворювати дані під час занесення у БД даних ПІП. Внаслідок описаних факторів можуть виникати проблеми при уніфікації записів зведених баз даних, що були отримані, наприклад, після злиття БД декількох департаментів великих компаній.

Таким чином, актуальною у галузі опрацювання БД постає задача виправлення помилок у полях, що описують ПІП. Також слід розглянути можливість профілактики помилок у подальшому, під час заповнення бази даних.

У статті запропоновано підхід щодо вирішення поставленої задачі. Він базується на пошуку кодових відстаней між рядками [1], додатково зважених числовими показниками – вагами.

Постановка задачі

У статті розглядається підхід до виправлення полів типу «Прізвище Ім’я По-батькові», що відповідають наступним умовам:

1. Структура – рядок (**string**) вигляду:
    a. Прізвище Ім’я По-батькові;
    b. Прізвище Ім’я;
    c. Прізвище Ім’я+По-батькові (Ім’я і По-батькові злиті в одному слові);
    d. Прізвище І. П. (І. П. - ініціали);
2. Спотворення:
    a. Зміна декількох букв запису на довільну іншу;
    b. Видалення частини літер;
    c. Зрізання закінчень слів (у випадку 1.c зрізані можуть бути обидва слова);

Термінологія

Кодова відстань. У якості функції пошуку кодової відстані між рядками обрано *відстань Левенштейна (редакційна відстань)* [1], що, за визначенням, чисельно дорівнює найменшій кількості операцій вставки, видалення чи заміни символу. Математичний запис: $dist(\vec{v1},\vec{v2})$, $\vec{v1},\vec{v2}$ – рядки-значення, Dist – кодова відстань.

Еталонною множиною будемо називати множину значень (параметрів), які вважаються коректними. Математичний запис: $ES = \vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$, ES (etalon set) – еталонна множина, $\vec{v}_i$ – елемент еталоної множини (далі – *еталонне значення*).

Для програмного комплексу – системи виправлення помилок – еталонну множину будемо називати *еталонною структурою даних*, а еталонне значення – *еталонним записом*.

Зважена еталонна множина – еталонна множина, кожному значенню якої поставлено у відповідність числовий показник, котрий додатково характеризує його. Цей показник назвемо *вагою*. Математичний запис: $ESw = \vec{V}_1, \vec{V}_2, \dots, \vec{V}_n$, ESw (etalon set with weight) – зважена еталонна множина, $\vec{V}_i = \{\vec{v}, w\}$ – елемент зваженої еталонної множини (далі – *зважене еталонне значення*), для якого: $\vec{v}$ – значення, w – вага.

Опис алгоритму

Виправлення імен. Виправлення імені виконується на основі використання зважених еталонних структур даних чоловічих і жіночих імен. Розглянемо алгоритм:

1. Знайти найближчі зважені значеня для зважених еталонних структур даних: окремо чоловічих і жіночих імен.
2. Вибрати з отриманих значень найближче до імені, що виправляється. Зберегти його як результат.

Перевірка статі. Перевірку статі виконаємо на основі наступних зважених критеріїв:

1. Для прізвища: закінчення на «а» - критерій належності до жіночої статі. Винятки: прізвища, що не змінюються в залежності від статі: Гук, Плющ, Жук, тощо;
2. Для ім'я: порівнюється кодова відстань між ближчими записами з еталонних структур даних чоловічих і жіночих імен;
3. Для ім'я по-батькові: закінчення на «-на» / «-ич».

Утворення імені по-батькові з чоловічого імені. Як матеріал, на основі якого була реалізована функція утворення імені по-батькові, було обрано наступні науково-популярні статті [2, 3], які охоплюють найбільш поширені випадки подібного словотворення.

Розділ злитих слів Ім'я+По-батькові. Авторами пропонується провести розбиття з використанням *роздільного індексу*. Алгоритм є ітеративним. Розглянемо його опис:

1. Встановивши значення роздільного індексу рівним:
    a. довжині слова;
    b. індексу першого символу.
2. Взяти підрядок зі слова, що розділяється, у проміжку:
    a. з початку слова, до роздільного індексу;
    b. з роздільного індексу, до кінця слова.
2. На основі еталоної структури даних провести пошук найближчого корректного імені для отриманої підрядку:
    a. використовуючи еталоні структури данних чоловічих і жіночих імен;
    b. використовуючі еталону структуру данних лише чоловічих імен.
3. Зберегти кодову відстань;
4. Виконати:
    a. декримент роздільного індексу;
    b. інкримент роздільного індексу.

5. Проводити ітерації доки кодова відстань до чергового слова не стане меншою за відстань, збережену на попередній ітерації, або схожість виявляється меншою за деяку граничну константу, або довжина досліджуваного підрядка менша за деяке граничне число.

Пункти *a)* вказані для алгоритму відокремлення імені, пункти *b)* – для імені по-батькові.

На рис. 1 представлена ілюстрація до описаного алгоритму (на етапі виправлення імені).

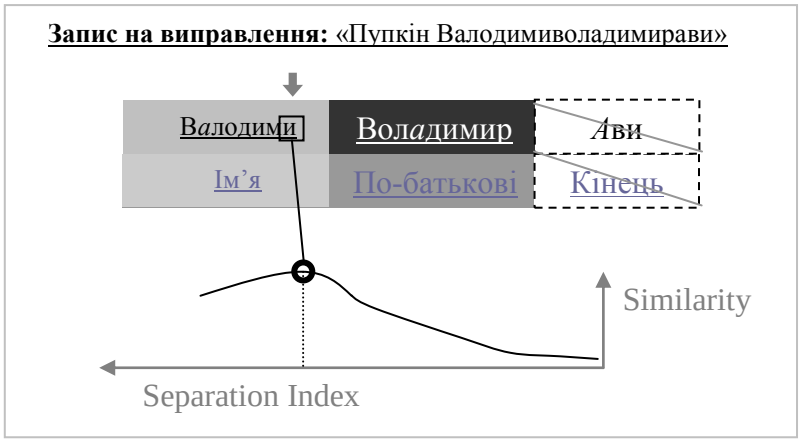

Рис. 1. Розділ злитих слів Ім'я+По-батькові

Реалізація адаптивної системи. На основі алгоритму можна організувати напівавтоматичну обробку записів ПІП. При цьому на кожному етапі виправлення программа шукає в еталонних структурах данних не одне найближче значення, а усі значення, що ближчі за деяку визначенну мінімальну відстань. Якщо кодова відстань для ближчого значення з отриманого списку виявляється меншою за визначенну константу критичної точності, алгоритм самостійно виконує виправлення. Інакше – зберігає варіанти виправлення, і, після опрацювання всіх записів, пропонує користувачеві самостійно вибрати найбільш корректний варіант зі створеного таким чином переліку.

При описаній напіавтоматичній обробці вибір варіанту виправлення користувачем та автоматичний вибір (вибір за константою критичної точності) змінює вагу обраного для виправлення запису. Таким чином, при існуванні схожих імен у БД, алгоритм зможе більш точно робити вибір коректного виправлення. Чим більше записів буде оброблено, тим точніший результат опрацювання зможе надавати програма, і тим рідше буде виникати необхідність втручання користовуча для уточнення виправлення.

Висновки

Запропонований спосіб обробки спотворених даних полів типу «ПІП» з використанням еталонних зважених словників дає змогу найбільш точно виконати приведення до стандартного вигляду і навіть доповнити деяку відсутню інформацію.

Після певних налаштувань та вдосконалень, реалізація даного способу дає змогу автоматизувати процес злиття реальних баз даних платників послуг, що вимагав нормалізації та уніфікації полів «ПІП». Це підтверджує ефективність запропонованого алгоритму.

У подальшому планується додати елатонні структури даних для винятків – для більш коректного формування імені по-батькові. Користувач сам зможе задавати слова на виправлення. Аналогічно можна реалізувати доповнення еталонних структур даних новими значеннями.

Література

1. Расстояние Левенштейна [Електронний ресурс http://ru.wikipedia.org/wiki/Расстояние_Левенштейна], дата візиту 27.10.2010
2. О построении словаря / Об отчествах [Електронний ресурс / http://slovari.donpac.ru/lang/ru/ivoc/name/name_postr.html], дата візиту 27.10.2010
3. Народная этимология / Образование и написание отчеств [Електронний ресурс http://fictionbook.ru/author/anatoliyi_pashalov/udivitelnaya_yetimologiya/read_online.html?page=3], дата візиту 27.10.2010