УДК 004.891

к.т.н., доц. **Джулій В.М.** (ХмНУ)
Лукіна К.В. (ВІТІ)
Солодєєва Л.В. (ВІКНУ)
Хлистун І.А. (ХмНУ)

АЛГОРИТМ ПРИЙНЯТТЯ РІШЕННЯ ІДЕНТИФІКАЦІЇ ФІЗИЧНИХ ОСІБ НА ОСНОВІ СИСТЕМИ ПРАВИЛ І ВАГ

Проведений аналіз напрямків розвитку сучасних баз даних показує, що склалися і формуються за останні роки тенденції розвитку інформаційних технологій істотно впливають на функціональні можливості автоматизованих систем. Задача встановлення відповідності між окремими об'єктами - побудова процедур ототожнення ускладнюється відсутністю серед загальних атрибутів відповідних один одному таблиць різних БД первинних ключів і наявністю помилок операторського введення. З урахуванням специфіки роботи з персональними даними пропонується вирішення наступних прикладних задач: повна ідентифікація клієнта при наявності спотворень інформації в базі даних або в пошукових запитах; усунення дублікатів записів при надходженні до БД з множинних джерел зі слабоструктурованою інформацією; пошук і коректування помилок в персональних даних клієнтів (фізичних і юридичних осіб). Укрупнений алгоритм даного підходу складається з трьох основних блоків: формування масиву «подібних» людей; використання не суворої відповідності серед масиву «подібних» людей; відпрацювання виняткових ситуацій. Дозволяє: виконувати функцію ідентифікації фізичної особи; при створенні реєстрів населення може допомогти при первинному об'єднанні накопичених відомчих БД; зберегти інформаційну цілісність, а також знизити зашумленість даних, обумовлену наявністю помилок операторського введення; виробляти об'єднання записів, відсоток схожості, по заданому набору полів яких вище встановленої межі.

Алгоритм ідентифікації фізичних осіб та алгоритм не суворого порівняння рядків, дозволяють оцінити ступінь схожості даних клієнтів. Розроблена система правил і ваг є основою для прийняття рішення по ідентифікації фізичних осіб.

На основі запропонованих алгоритмів розроблений програмний модуль, який призначений для пошуку та усунення дублікатів записів в базі даних за допомогою операції не суворої відповідності та інтегрується із засобами СУБД.

Ключові слова: база даних, нечіткий пошук, порівняння рядків, пошук даних, алгоритми, інформаційна система.

Вступ. Неухильне зростання обсягів даних викликає необхідність широкого використання передових інформаційних технологій для ефективного управління потоками

даних. При цьому, найбільшу значимість набувають завдання створення ефективних інструментів оцінки та контролю зростаючих потоків інформації, оптимізацій процедур обробки, агрегації, узагальнення, пошуку та аналізу даних. Зростає попит на створення, як корпоративних автоматизованих інформаційних систем, так і окремих спеціалізованих рішень. Автоматизовані інформаційні системи (АІС) розробляються на основі інформаційно-аналітичних баз даних, які використовуються в якості ключового елемента системи і забезпечують зберігання і обробку всієї сукупності даних, що надходять від підрозділів і філій. З точки зору технологій, АІС представляє набір апаратних засобів, технологій, методів і алгоритмів, спрямованих на підтримку життєвого циклу інформації і включає три основні процеси: обробку даних, управління інформацією та управління знаннями.

Для багатьох організацій інформація є основним активом. Спотворення або пошкодження важливої інформації може призвести до суттєвих фінансових втрат і репутаційним ризикам. На основі проведеного аналізу даних, що отримані з відкритих джерел і наукових публікацій, можна виділити основні види втрат, що виникають внаслідок помилок і спотворень інформації в базах даних. Втрати внаслідок невірного, не якісного надання послуг («брак» в інформації). В середньому організація втрачає 25-40% часу співробітників, від втрат данного виду. Втрати оплачуваного часу співробітників на непродуктивну діяльність. Даний вид втрат може досягати, наприклад, у менеджерів середньої ланки більше 50% робочого часу, у менеджерів низової категорії до 80%. Втрати внаслідок використання не оптимальних технологічних ланцюжків. За цими причинами в середньому організація втрачає близько 35% робочого часу задіяних співробітників і це може призвести до подорожчання однієї операції до 100%. Втрати часу, грошових коштів, клієнтів по причині відсутності або дублюванні інформації. Втрати становлять близько 15% часу співробітників, що спричиняє збільшення вартості виконуваної операції.

Основним чинником, що стимулює розвиток технологій пошуку, є поява великої кількості електронних бібліотек і архівів, що містять значні обсяги актуальних знань. Продуктивність і ефективність будь-якої системи зберігання інформації безпосередньо залежить від ефективності та продуктивності пошукових систем. Саме пошукова система визначає, чи перетворяться в знання численні розрізнені дані, що надходять по різних каналах зв'язку і накопичуються в різноманітних базах даних та електронних архівах.

Найбільш поширеним видом інформаційних ресурсів для організацій, що працюють з персональними даними (бюро кредитних історій, банки, страхові організації, будь-які організації з досить крупним штатом співробітників) є тексти на природних мовах. Цим обумовлено широке застосування в таких системах технологій текстового пошуку. Дані технології використовуються не тільки в системах, побудованих за принципом традиційних текстових систем, але і для пошуку в колекціях, організованих у вигляді веб-сайтів, а також для пошуку в глобальної мережі Інтернет. Управління дисковим простором, збільшення ємності систем зберігання і зростання їх продуктивності, міграція даних із застарілих сховищ на нові - всі ці, і багато інших завдань доводиться постійно вирішувати компаніям, що використовують системи зберігання даних.

Постановка задачі. Разом з тим, існують фактори, що стримують розвиток АІС. При організації пошуку в базах персональних даних клієнтів виникають характерні проблеми, пов'язані з наявністю в запитах орфографічних і фонетичних помилок, помилок введення інформації, а також відсутністю єдиних стандартів транскрипції з іноземних мов. Внаслідок цього задача пошуку в базах персональних даних не може бути повною мірою вирішена тільки методами перевірки на точну відповідність. Стає актуальною задача розробки спеціальних методів і технологій текстового пошуку з використанням нетривіальних рішень, в тому числі на основі операцій не суворої відповідності. Універсальної методики пошуку в умовах зашумленості даних не існує, оскільки кожна проблема має власну оригінальну специфіку. Для вирішення виниклих проблем, необхідно використовувати алгоритми які здатні відшукати всі лексикографічно близькі до шаблону пошуку слова, що відрізняються замінами, пропусками і вставками символів. Таким чином, автоматично стає допустимою

помилка, як у вхідних даних, так і в термінах запиту. В даний час можливості виконання пошуку за подібністю не використовуються в СУБД. Таким чином, виникає задача розробки алгоритмів виконання спеціальних реляційних операцій, що виникають в задачі ототожнення записів. Проведений аналіз напрямків розвитку сучасних баз даних показує, що склалися і формуються за останні роки тенденції розвитку інформаційних технологій істотно впливають, у тому числі і на функціональні можливості автоматизованих систем. Задача встановлення відповідності між окремими об'єктами - побудова процедур ототожнення в даний час не має задовільного рішення. Існуючі роботи, присвячені інтеграції БД, дозволяють здійснити тільки інтеграцію схем БД, але не пропонують способів побудови процедур ототожнення. Побудова процедур ототожнення ускладнюється відсутністю серед загальних атрибутів відповідних один одному таблиць різних БД первинних ключів і наявністю помилок операторського введення.

З урахуванням специфіки роботи з персональними даними пропонується вирішення наступних прикладних задач: повна ідентифікація клієнта при наявності спотворень інформації в базі даних або в пошукових запитах; усунення дублікатів записів при надходженні до БД з множинних джерел зі слабоструктурованою інформацією; пошук і коректування помилок в персональних даних клієнтів (фізичних і юридичних осіб).

Основна частина. Однією з задач при обміні інформацією про клієнта, є його однозначна ідентифікація. Одним, з таких рішень є ідентифікація фізичних осіб шляхом порівняння їх основних реквізитів. Таке рішення не завжди буде прийнятне при використанні, простого порівняння реквізитів, тому що по ряду причин, реквізити однієї і тієї ж особи, взяті з двох різних БД можуть не співпадати, тому що вони не завжди доступні (наприклад, через помилки в джерелі даних); присвоєні не всім категоріям громадян (ідентифікаційний код, страхове свідоцтво пенсійного фонду); реквізити документа змінені. внаслідок втрати/псування/за бажанням громадянина; реквізити документа відрізняються внаслідок помилки при занесенні в БД. Готової методики по такому виду ідентифікації на даний час не існує. На основі проведених досліджень і експериментів було знайдено рішення, що дозволяє проводити ідентифікацію фізичних осіб в БД з максимальною точністю. В результаті була розроблена технологія, із застосуванням якої може бути організований більш ефективний інформаційний обмін. Укрупнений алгоритм даного підходу складається з трьох основних блоків: формування масиву «подібних» людей; використання не суворої відповідності серед масиву «подібних» людей; відпрацювання виняткових ситуацій.

Розглянемо основні поняття що використовуються даним алгоритмом.

Функція релевантності - реалізована на основі алгоритму порівняння підрядків і визначає близькість рядкових значень у відсотках. Аргументами функції є два рядки і параметр порівняння (N), що представляє із себе максимальну довжину підрядків, що беруть участь в порівнянні. Як результат функція повертає відсоток релевантності, де 0% вказує на абсолютну розбіжність рядків, а 100% на тотожну рівність. Порівняння відбувається за наступним алгоритмом: функція створює два набори підрядків (довжина підрядків обмежується параметром порівняння). Для підрядків однакової довжини в двох наборах функція знаходить підрядки першого рядка, які є в другому підрядку, додає кількість співпадань підрядків другого рядка з підрядками першого. Відношення суми співпадань до числа варіантів, приведене до процентного виду, запам'ятовується як проміжний коефіцієнт релевантності для даної довжини підрядків, далі береться середнє значення всіх проміжних коефіцієнтів і повертається функцією як відсоток релевантності вхідних рядків.

Функцію релевантності від двох рядків S_{t1} і S_{t2} довжиною l_1 і l_2 відповідно і максимальної довжини підрядків N визначимо наступним чином:

1. Формуємо набори всіх можливих підрядків довжиною до N:

$$G_j(i) = \{g_{j1}(i), \ldots, g_{jk}(i), \ldots, g_{jn}(i)\}; \quad j = 1,2; \quad i = \overline{1,N}; \quad n = l_j - i + 1; \qquad (1)$$

де *i*- довжина підрядка; *j* - номер вхідного рядка; *n* - кількість підрядків довжиною *i* в *j*-му слові.

2. Кожному набору $G_j(i)$ поставимо у відповідність множину $G_j^*(i)$, елементи яких не повторюються із набору $G_j(i)$, тобто повторюваним елементам набору $G_j(i)$ в множині $G_j^*(i)$ буде відповідати один елемент:

$$G_j(i)=\{g_{j1}(i),...,g_{jk}(i),...,g_{jn}(i)\};\quad j=1,2;\quad i=\overline{1,N};\quad n=l_j-i+1; \tag{2}$$

де *m* - кількість неповторюваних підрядків довжиною *i* в *j*-му слові.

3. Значення функції релевантності $FR=(l_1,l_2,N)$ обчислюється за наступною формулою:

$$FR=(l_1,l_2,N)=\frac{\sum_{i=1}^{N} fr(i)}{N}, \tag{3}$$

$$fr(i)=\frac{|{}^*G_1(i)|+|{}^*G_2(i)|}{|G_1(i)|+|G_2(i)|}, \tag{4}$$

$${}^*G_j(i)=G_j(i)\cap G_j^*(i), \tag{5}$$

де $g_j(i)\in {}^*G_j(i)\Rightarrow \exists\, g_k^*(i): g_j(i)=g_k^*(i)$ тобто набір ${}^*G_j(i)$ складається з елементів набору $G_j(i)$, для яких є рівні у множині $G_j^*(i)$. $G_j(i)$ - набір підрядків довжиною *i* рядка l_j; $|G_j(i)|$ - кількість елементів у наборі підрядків $G_j(i)$; $G_j^*(i)$ - множина, в якій не повторюються підрядки набору $G_j(i)$; $|G_j^*(i)|$ - кількість елементів у наборі підрядків $G_j^*(i)$; $|{}^*G_j(i)|$ - кількість елементів у наборі підрядків ${}^*G_j(i)$; N - максимальна довжина підрядка.

Вага - умовний коефіцієнт реквізиту. Він залежить від повноти, достовірності, і актуальності реквізиту. Вага визначає значимість реквізиту для ідентифікації

Правило - поєднання реквізитів клієнта, за якими здійснюється пошук. Механізм пошуку за правилами такий, що при пошуку клієнта порівнюються тільки ті реквізити, які вказані в правилах. Наприклад, при використанні правила 1 (табл. 1) порівнюються тільки «Прізвище», «Ім'я» і «Дата народження», при цьому не враховуються інші реквізити. Для кожного правила визначається його сумарна «вага», яка складається з суми «ваг» реквізитів, що входять у дане правило (див. табл.).

Таблиця

Ваги правил

№ правила	Реквізити	Сумарна вага реквізитів	Поріг ідентифікації по правилу
Правило 1	Прізвище (5) + Ім'я (4) + По батькові (3)	12	11
Правило 2	Прізвище (5) + Ім'я (4) + Дата народження (4)	13	12
Правило 3	Ім'я (4) + По батькові (3)+ Дата народження (4)	11	10
Правило 4	Дата народження (4) + По батькові (3) + Місце народження (3)	10	9

Розглянемо перший блок алгоритму – формування масиву «подібних» людей, який наповнюється з використанням правил. Для вибору єдиної вірної людини з масиву «подібних» людей, встановлюється поріг ідентифікації. Поріг ідентифікації необхідний для того, щоб виключити людину, яка не задовольняє умовам. Поріг ідентифікації встановлюється, як показано в табл. 1. Якщо поріг подолали більше однієї особи, то автоматизовано ідентифікувати особу неможливо. Така ситуація відпрацьовується оператором. Наступним кроком технології є вибір людини із застосуванням функції релевантності.

Основою роботи алгоритму ідентифікації фізичних осіб (ФО) є умова:

$$\sum_{i=1}^{n} p_j(i) \cdot (R_i \cdot w_i + L_i) \geq k_j; \qquad j = \overline{1,m}; \qquad i = \overline{1,n}; \tag{6}$$

де $p_j(i)$ - елемент правила ідентифікації. Правило – поєднання реквізитів ФО, за якими відбувається порівняння, $p_j(i) = 1$, якщо *i*-ий реквізит входить в *j*-е правило, $p_j(i) = 0$, якщо не входить. R_i - результат роботи функції релевантності від *i*-их реквізитів; w_i - вага реквізиту. Залежить від повноти, достовірності та актуальності реквізиту. Визначає значимість реквізиту для ідентифікації ФО; L_i - підвищувальний коефіцієнт розрахований на підставі відстані Левенштейна між *i* -ми реквізитами; k_j - поріг ідентифікації правила. Необхідний для виключення записів, які не пройшли ідентифікацію за правилами. Якщо поріг пройшли декілька записів, то автоматизовано ідентифікувати особу неможливо. Така ситуація відпрацьовується оператором; *m* - кількість правил; *n* - кількість реквізитів, що беруть участь в порівнянні. При цьому, якщо (6) вірно хоча б для одного *j*, то реквізити ФО пройшли ідентифікацію за правилом *j* і вважаються подібними. Дані записи надходять на розгляд аналітику, в іншому випадку записи вважаються різними і порівняння триває. Для визначення кількості помилок, що усуваються із застосуванням не суворої відповідності проведено розрахунок:

$$P(A) = \frac{m}{n}, \tag{7}$$

де *m* - кількість реквізитів з помилками, *n* - загальне число реквізитів, *P* (*A*) - частота появи помилки в реквізиті.

Для отримання ймовірності появи рядка з помилкою в хоча б одному реквізиті використаємо суми ймовірностей:

$$P(A + B) = P(A) + P(B). \tag{8}$$

Ймовірність появи рядка з помилкою в хоча б одному реквізиті буде дорівнює сумі трьох ймовірностей. Отриманий експериментально відсоток помилок може варіюватися в обидві сторони і в широких межах, але на практиці не було відмічено випадків, коли в імпортованій з різних баз даних - джерел інформації відсутні помилки. Це пов'язано з наявністю людського фактора при обробці великих масивів даних по фізичним особам: друкарські помилки, помилки введення, нечіткі копії первинних документів та інші випадки.

Можливі результати роботи алгоритму: людина знайдена (це точно вона - співпали всі правила); людини не знайдено (її точно немає); знайдено кілька схожих людей, автоматизовано визначити не можливо, вимагає відпрацювання оператором.

Ситуації, які вимагають доопрацювання оператором, потрапляють в лог прийняття рішень. Відпрацювання логу здійснюється шляхом повторного звернення до джерела даних, первинних документів, особової справи і т.д.

Запропонований алгоритм дозволяє: виконувати функцію ідентифікації фізичної особи, при створенні реєстрів населення може допомогти при первинному об'єднанні накопичених відомчих БД; зберегти інформаційну цілісність, а також знизити зашумленість даних, обумовлену наявністю помилок операторського введення; виробляти об'єднання записів, відсоток схожості, по заданому набору полів яких вище встановленої межі.

Задачу пошуку за окремими атрибутами необхідно розбити на 3 підзадачі:

1. Пошук по рядках довжиною більше 50 символів, такі як: нотатки та коментарі інспектора, співробітників CALL-центрів, служб збору, що раніше працювали з клієнтом і т.п.

2. Пошук по відносно коротким полях з середньою довжиною менше 50 символів: назви місць роботи, назви вулиць, торговельних точок, юридичних осіб, телефонів, факсів, електронних адрес.

3. Пошук за прізвищами при ручному введенні реквізитів клієнтів або отриманням «на льоту» списку відповідних записів про клієнтів кредитним інспектором.

Спеціально розроблений алгоритм наближеного пошуку застосовується до першої задачі. Алгоритм пошуку за окремими атрибутами (рис. 1) полягає в наступному. Якщо дані про клієнта вводяться безпосередньо співробітником і включений блок не суворого пошуку, то поки вводяться інші дані, блок Metaphone аналізує прізвище, введене в спеціальне поле, і формується попередній масив подібних прізвищ клієнтів існуючих в базі даних. По мірі заповнення форми інформації про клієнта, даний масив аналізується з використанням функції релевантності та урізається шляхом відкидання записів, що не пройшли поріг ідентифікації. Після цього оператору виводяться записи, що залишилися, відсортованими в порядку убування відстані між знайденими прізвищами. Алгоритм пошуку за окремими атрибутами представлена на рис. 1.

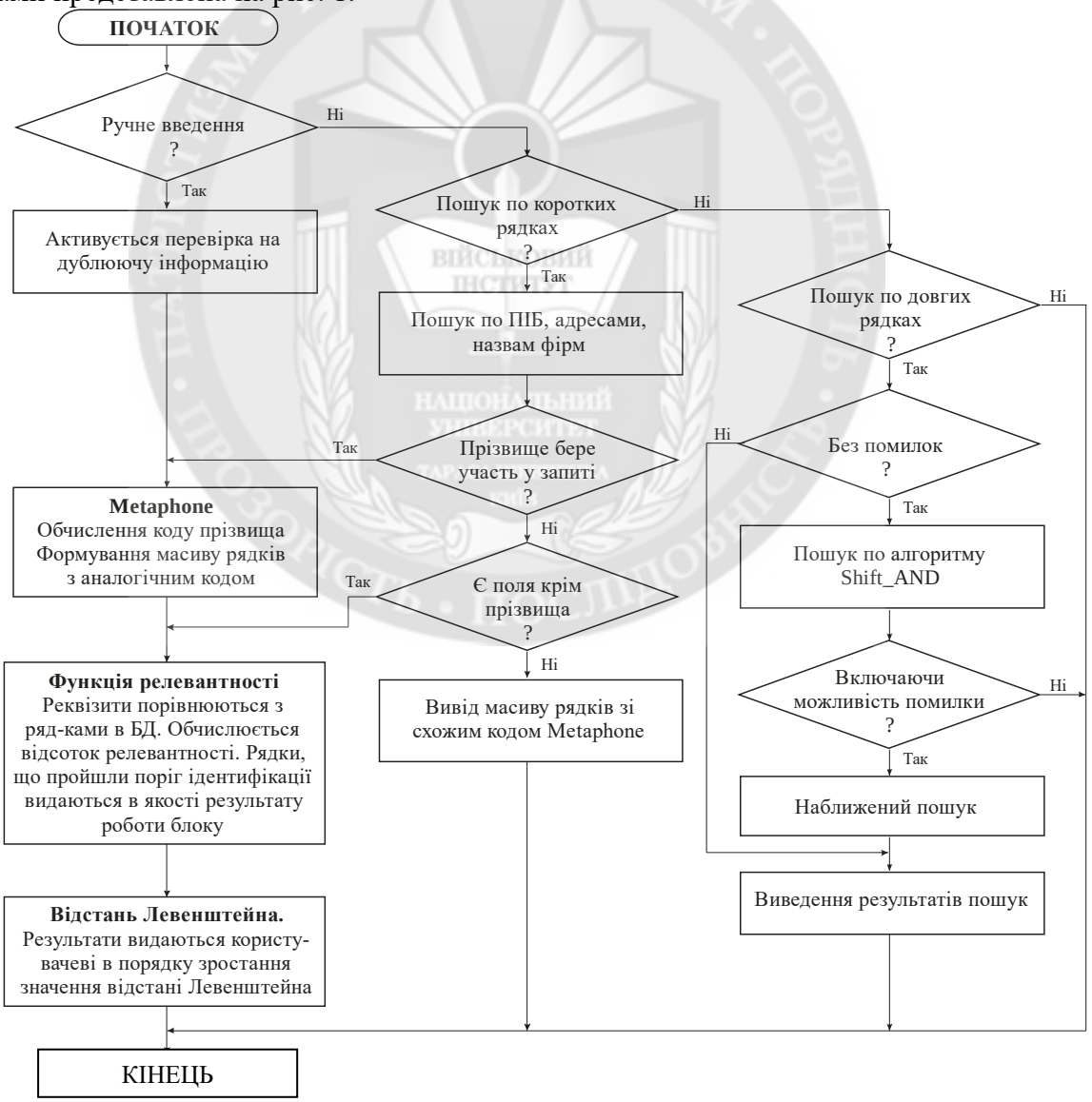

Рис. 1. Алгоритм пошуку за окремими атрибутами

Якщо здійснюється пошук за реквізитами і включений блок не суворого пошуку, то спочатку визначається, чи бере прізвище участь в запиті, якщо так, то використовуючи алгоритм Metaphone дане прізвище перетворюється і здійснюється попередній пошук в тому модулі, куди адресований запит. Далі результати цього попереднього пошуку аналізуються із застосуванням інших рядків запиту і функції релевантності, поступово зменшуючи результати вибірки. Якщо на прізвищі запит закінчується, і результати цього запиту містять менше 80 рядків, то результати видаються користувачеві, також відсортовані в порядку убування по відстані Левенштейна. Якщо ж прізвище не бере участі в запиті, то пошук відразу починається з обчислення функції релевантності. Якщо відбувається пошук ключових слів по текстових полях, то в разі включеного блоку не суворого пошуку використовується алгоритм наближеного пошуку, і алгоритм Shift-and в разі не включеного.

Запропонований алгоритм ототожнення об'єктів і усунення дублювання інформації, відрізняється побудовою функції релевантності, дозволяє знаходити рішення в найбільш загальному вигляді. Алгоритм дозволяє виробляти об'єднання записів, визначити відсоток схожості по заданому набору полів, яких знаходиться в встановлених межах.

Алгоритм ідентифікації фізичних осіб з використанням правил ідентифікації та алгоритму не суворого порівняння рядків, дозволяє оцінити ступінь схожості даних клієнтів. Розроблена система правил і ваг є основою для прийняття рішення в ідентифікації клієнтів. Запропонований алгоритм пошуку по атрибутах на основі функції релевантності, алгоритму фонетичної схожості, відстані Левенштейна і наближеного пошуку на базі модифікації алгоритму прямого перебору «Shift - AND».

На основі запропонованих алгоритмів створений програмний модуль «Автоматизація пошуку дублікатів в базі даних». Модуль призначений для пошуку та усунення дублікатів записів в базі даних за допомогою операції не суворої відповідності. На стороні сервера інсталюється серверна частина модуля із зазначенням робітників і проміжних баз даних. Зазначені бази даних доповнюються пакетами процедур і функцій, необхідних для роботи модуля. Структура програмного модуля «Автоматизація пошуку дублікатів в базі даних» показана на рис. 2. Інформація надходить з комерційних організацій - інформаційних партнерів, з державних органів. Отримані дані проходять попередню очистку, приводяться до потрібного формату і акумулюються в проміжній базі даних. Наступним кроком стає поповнення основних баз за допомогою розробленого програмного модуля.

Програмний модуль складається з наступних основних блоків:

1. Блок розпаралелювання потоків інформації. Для прискорення процесу поповнення баз даних, було вирішено організувати роботу модуля в декілька потоків.

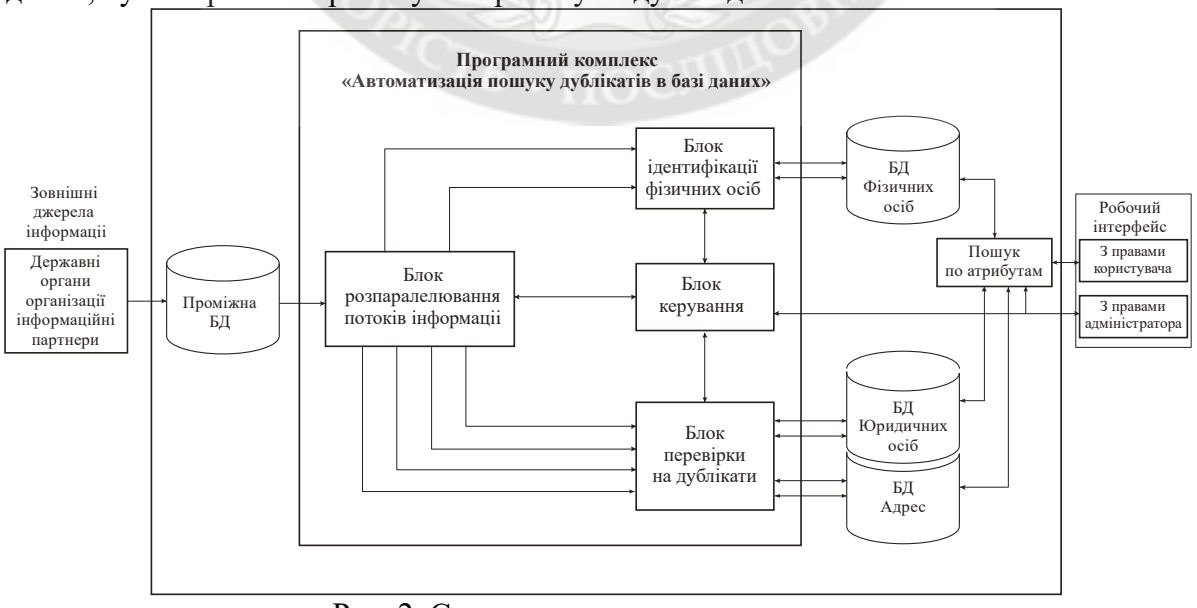

Рис. 2. Структура програмного модуля
«Автоматизація пошуку дублікатів в базі даних»

2. Блок перевірки на дублікати. Пакет процедур і функцій, що викликаються з блоку розпаралелювання стільки разів, скільки потоків даних було сформовано. Результатом роботи блоку - записи або визнаються новими і тоді додаються в необхідну базу даних, або визнаються дублікатом і вносяться з відповідною позначкою в лог виявлених дублікатів, якщо рішення про додавання /видалення прийняти не вдалося, тоді даний запис потрапляє в лог прийняття рішення і зберігається там до обробки аналітиком-адміністратором.

3. Блок ідентифікації фізичних осіб. Робота блоку ідентифікації фізичних осіб в цілому схожа на роботу блоку перевірки на дублікати, за винятком відмінностей у роботі алгоритму ідентифікації.

4. Блок керування. За допомогою інтерфейсу адміністратора, через блок керування можна в широких межах налаштувати, роботу модуля. У блоці розпаралелювання потоків інформації можна налаштувати кількість потоків, а також кількість записів у потоці. Для блоку ідентифікації фізичних осіб і блоку перевірки на дублікати блок керування служить джерелом порогових коефіцієнтів для роботи алгоритмів ідентифікації, а також сховищем логів прийняття рішення, логу виявлених дублікатів та загального логу роботи модуля. Крім цього блок керування містить процедури і функції, що відповідають за логіку обробки аналітиком-адміністратором виключень, що містяться в лозі прийняття рішень.

5. Блок пошуку по атрибутам. Блок розміщений осібно від основного модуля і встановлюється при необхідності в якості надбудови. Блок використовується в якості пошукової системи по базі даних.

Висновки. Алгоритм ідентифікації фізичних осіб з використанням правил ідентифікації та алгоритму не суворого порівняння рядків, дозволяє оцінити ступінь схожості даних клієнтів. Розроблена система правил і ваг є основою для прийняття рішення ідентифікації фізичних осіб. Запропонований алгоритм пошуку по атрибутах на основі функції релевантності, алгоритму фонетичної схожості, відстані Левенштейна і наближеного пошуку на базі модифікації алгоритму прямого перебору «Shift - AND».

Запропонований алгоритм дозволяє: зберегти інформаційну цілісність, а також знизити зашумленість даних, обумовлену наявністю помилок операторського введення; виробляти об'єднання записів, в яких відсоток схожості по заданому набору полів вище встановленої межі; виробляти усунення дублювань як на підставі автоматично налаштованих правил (автоматичний режим), так і з втручанням людини в особливо складних випадках (ручний режим). На основі запропонованих алгоритмів розроблений програмний модуль «Автоматизація пошуку дублікатів в базі даних». Модуль призначений для пошуку та усунення дублікатів записів в базі даних за допомогою операції не суворої відповідності. На стороні сервера інсталюється серверна частина модуля із зазначенням робітників і проміжних баз даних.

ЛІТЕРАТУРА:

1. Ахо А. Структуры данных и алгоритмы./ А. Ахо, Д. Хопкрофт, Д.Ульман. – М.: Вильяме, 2009. – 400 с.

2.Вирт Н. Алгоритмы и структуры даннях / Н. Вирт. – М.: ДМК Пресс, 2010. – 272 с.

3. Гагарина Л.Г. Разработка и эксплуатация автоматизированных информационных систем: учеб. пособие / Л.Г. Гагарина, Д.В. Киселев, Е.Л. Федотова. – М.: ИД «Форум»: Инфа-М, 2007. – 384 с.

4. Гагарина Л.Г. Алгоритмы и структуры данных. / Л.Г. Гагарина, В.Д. Колдаев. – М.:Инфра-М, 2009. – 304 с.

5. Гайдамакин Н.А. Автоматизированные информационные системы, базы и банки данных / Н.А. Гайдамакин. – Москва «Гелиос АРВ», 2002. –368 с.

6. Кнут Д.Э. Искусство программирования / Кнут Д.Э. –Том 4. Выпуск 2. Генерация всех кортежей и перестановок. – Ml:: Вильяме, 2008. – 160 с.

7. Макленнен Д. Microsoft SQL Server 2008 / Макленнен Д., Танг Ч., Криват Б. Data Mining - интеллектуальный анализ данных. – СПб.: БХВ-Петербург, 2010. – 700 с.

REFERENCES:

1 Aho Data Structures and Algorithms. / A. Aho, J. Hopcroft, D..Ulman // - M .: Williams, 2009, - 400.

2.Virt N. Structures and Algorithms dannyah / Wirth // - M .: DMK Press, 2010 - 272 p.

3. Gagarin LG / design and operation of automated information systems / LG Gagarin, DV, Kiselev E.L.Fedotova //: Proc. allowance. M .: ID "Forum": Infa-M, 2007. 384 p.

4. Gagarin LG Algorithms and Data Structures. /L.G. Gagarin,VD Koldaev// - Moscow: Infra-M, 2009.- 304.

5. Gaydamakin NA Automated information systems, databases and data banks. / NA Gaydamakin // Moscow "Helios ARV" 2002. 368 p.

6. DE Knuth Art of Computer Programming. / DE Knuth // Volume 4 Issue 2 Generating all tuples and permutations. - Ml :: Williams, 2008 - 160 p.

7. Macclenny D. Microsoft SQL Server 2008 / Macclenny, D., Tang, C. Krivat B .// Data Mining - Data Mining - St. Petersburg .: BHV-Petersburg, 2010 - 700 p.

Рецензент: д.т.н., проф. Лєнков С.В., начальник науково-дослідного центру Військового інституту Київського національного університету імені Тараса Шевченка

к.т.н., доц. Джулий В.М., Лукина Е.В., Солодеева Л.В., Хлистун И.А.

АЛГОРИТМ ПРИНЯТИЯ РЕШЕНИЯ ИДЕНТИФИКАЦИИ ФИЗИЧЕСКИХ ЛИЦ НА ОСНОВЕ СИСТЕМЫ ПРАВИЛ И ВЕСОВ

Проведенный анализ направлений развития современных баз данных показывает, что сложились и формируются за последние годы тенденции развития информационных технологий существенно влияют на функциональные возможности автоматизированных систем. Задача установления соответствия между отдельными объектами - построение процедур отождествления осложняется отсутствием среди общих атрибутов соответствующих друг другу таблиц различных БД первичных ключей и наличием ошибок операторского ввода. С учетом специфики работы с персональными данными предлагается решение следующих прикладных задач: полная идентификация клиента при наличии искажений информации в базе данных или в поисковых запросах; устранения дубликатов записей при поступлении в БД из множественных источников с слабоструктурированные информацией; поиск и корректировка ошибок в персональных данных клиентов (физических и юридических лиц). Укрупненный алгоритм данного подхода состоит из трех основных блоков: формирование массива «подобных» людей; использование не строгого соответствия среди массива «подобных» людей; отработка исключительных ситуаций. Позволяет: выполнять функцию идентификации физического лица; при создании реестров населения может помочь при первичном объединении накопленных ведомственных БД; сохранить информационную целостность, а также снизить зашумленность данных, обусловленную наличием ошибок операторского ввода; производить объединение записей, процент схожести, по заданному набору полей которых выше установленного предела.

Алгоритм идентификации физических лиц и алгоритм нестрогого сравнения строк, позволяют оценить степень сходства данных клиентов. Разработана система правил и весов является основой для принятия решения по идентификации физических лиц.

На основе предложенных алгоритмов разработан программный модуль, который предназначен для поиска и устранения дубликатов записей в базе данных с помощью операции не строгого соответствия и интегрируется со средствами СУБД.

Ключевые слова: база данных, нечеткий поиск, сравнение строк, поиск данных, алгоритмы, информационная система.

Ph.D. Dzhuliy V.M., Lukina E.V., Solodeeva L.V., Hlistun I.A.

HE DECISION-MAKING ALGORITHM OF THE NATURAL PERSON IDENTIFICATION BASED ON THE SYSTEM OF RULES AND WEIGHT

The analysis of the modern databases directions development shows that tendencies of the information technologies development significantly affects the functionality of automated systems. The task of establishing a correspondence between the individual objects - the construction of identification procedures is complicated by the absence of the common attributes corresponding to each other tables of

different database primary key, and the presence of operator input errors. It offers a solution for the following applications taking into account the specificity of personal data: full identification of the client in the presence of misstatements in the database or in search queries; eliminate duplicate records when entering the database from multiple sources with semistructured information; Search and correction of errors in the personal data of customers (individuals and legal entities). The integrated algorithm of this approach consists of three main blocks: the formation of an array of "similar" people; Use no strict correspondence among an array of "similar" people; working out exceptions. It allows you to: perform the function of identifying a natural person; in creating public registries can help during the initial merger accumulated departmental databases; preserve the integrity of the information, as well as to reduce the noise in the data, due to the presence of operator input errors; make association records, the percentage of similarity, given a set of fields which are higher than the set limit.

Algorithm for identification of individuals and the algorithm lax string comparison, allow us to estimate the degree of similarity of customer data. A system of rules and the balance is the basis for a decision on the identification of individuals.

The software module is developed on the base of the basis of the proposed algorithms which is designed to find and remove duplicate records in the database via the operation is not strict compliance and integrates with database tools.

Keywords: database, fuzzy searches, a string comparison, data mining, algorithms, information system.