ВСТАНОВЛЕННЯ ЗВ'ЯЗКІВ МІЖ НАБОРАМИ ДАНИХ ПРИ СТВОРЕННІ СХОВИЩА ДАНИХ

канд. техн. наук, проф. О.Б. Кунгурцев,
канд. техн. наук О. А. Блажко, С. В. Ковальчук, М. О. Скрипкін

Одеський національний політехнічний університет, Україна

Розглядається процес створення сховища даних електронних документів Веб-порталу відкритих даних, для скорочення трудомісткості якого запропоновано розвиток методу порівняння текстів шляхом визначення інтегральної близькості структурованих текстів та їх елементів у вигляді рядків і стовпців, що дозволяє автоматизувати процес встановлення зв'язку між наборами при створенні сховища. Програмне забезпечення методу апробовано на документах з сайту головного управління статистики в Одеській області та громадського веб-порталу відкритих даних Одеській області.

Вступ: Відомо, що сховища даних створюються з урахуванням таких особливостей [1]: джерел даних у формі традиційних систем реєстрації операцій, електронних документів або наборів даних, а також операцій з даними на рівні вилучення, перетворення, завантаження, аналізу і представлень результатів аналізу. Якщо в державних установах існує система електронного документообігу, яка зберігає табличні дані у файлах форматів текстових процесорів, наприклад, *RTF/DOC(X)/ODT*, або електронних таблиць, наприклад, *XLS(X)/ODS*, тоді існує і можливість їх окремого зберігання у вигляді табличних наборів даних. Постанова Кабінету Міністрів України № 835 від 21.10.2015 «Про затвердження Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних» [2] зобов`язала державні установи розміщувати публічні дані на національному порталі за адресою http://data.gov.ua в структурованих текстових форматах *CSV/XML/JSON*. Обробку 7933 наборів даних, які містять посилання на 19082 файли різних форматів, проведена в роботі [3], показала, що: *RTF/DOC(X)/ODT* – 33%, *XLS(X)/ODS* – 23%, *PDF* – 15%, *CSV* – 11%, *XML* – 6%. В результаті структурного аналізу документів *RTF/DOC(X)/ODT*-форматів було визначено, що 60% з них містять таблиці. Ця статистика, з одного боку, вказує на значну трудомісткість ручного процесу перетворення даних з документів офісних систем у *CSV*-формат (*CSV*-таблицю), а з іншого боку, вказує на відсутність автоматизованого процесу створення сховища даних з наборами даних, яке зв`язує набори даних за семантичними зв`язками між наборами даних з метою отримання нової інформації та інтелектуальної аналітичної обробки.

Метою роботи є автоматизація процесу встановлення зв'язку між наборами даних з CSV-таблиць при створенні сховища даних для підвищення ефективності їх подальшої аналітичної обробки.

Для досягнення мети вирішено наступні задачі:

- виділення в наборах даних термінів;
- визначення близькості таблиць *CSV*-таблиць по кількості однакових термінів;
- виявлення в *CSV*-таблицях подібних рядків і стовпців.

Виділення термінів з аналізованих документів. Термін – це слово, стійке словосполучення або скорочення, яке виражає і певною мірою класифікує в даній предметній області певне поняття чи сутність.

У даній роботі використані результати досліджень з автоматизованої побудови словників термінів виконані авторами для групи слов'янських мов [4, 5]. Результатом побудови словника термінів для документу T є текст *Tres*, який містить в собі всі терміни $term_i$ з вказаною кількістю повторень m_i та текстом терміну tx_i. Кожен запис $term_i$ має вигляд: $term_i =< tx_i, m_i >$.

Встановлення зв'язності документів на основі знайдених термінів.

Масив з усіма результатами *Tres* має вигляд: $AllTres = \{Tres_j\}\, j = 1, k$.

Для пошуку спільної кількості m_i термінів виконується порівняння термінів кожної пари документів *Tres*. У результаті знаходяться відсотки збігів термінів $proc_j$ і $proc_{j+1}$ для кожної пари документів.

Урахування ваги терміну. Чим більша кількість повторень терміну тим вища його «вага» у документі. Для обліку ваги терміна введено відношення m_z / N – кількості повторень унікального терміну відносно усіх знайдених термінів в $Tres_j$ та відношення $m_z / Nrez$ – кількості повторень унікального терміну відносно усіх спільних термінів між $Tres_j$ та $Tres_{j+1}$. Тоді відсоток близькості документів з урахуванням ваги термінів буде визначається как:

$$proc = (procTerm_z + procTerm_{z+1} ... procTerm_x) * kof , \quad (1)$$

де $kof = \sqrt{x}$ – коефіцієнт кількості спільних унікальних термінів.

Виявлення подібних стовпців і рядків в таблицях. Після виявлення близькості текстів по термінах здійснюється перевірка по стовпцях і рядках двох текстів з таблиць T_j і T_{j+1}. Для цього виявляються всі таблиці в цих документах і по кожній з них записуються усі заголовки рядків і стовпців.

Представимо результат аналізу тексту T у вигляді множини таблиць:

$$Ttable = \{table_y\} y = 1, u , \quad (2)$$

де кожна знайдена таблиця представляє собою кортеж у вигляді:

$$table_y =< rows_y, columns_y > , \quad (3)$$

де $rows_y$ – масив записів усіх заголовних рядків в таблиці $table_y$, а $columns_y$ – масив записів усіх заголовних стовпців в таблиці $table_y$. Після пошуку термінів в кожному заголовку рядків і стовпців формуються множини $row_q = \{tx_a\} a = 1, b$ та $column_e = \{tx_f\} f = 1, g$.

Усі таблиці $table_y$ двох порівнювальних документів T, порівнюються кожна з кожною по стовпцях і рядках. Відсоток близькості *p* визначається виразом:

$$p = com / mid * 100\% , \quad (4)$$

де *com* – кількість спільних термінів, а mid – середня кількість термінів.

Середній відсоток близькості *pMid* визначається виразом:

$$pMid = (\sum p) / \arg , \quad (5)$$

де *arg* – кількість не нульових порівнянь між заголовками рядків або стовпців.

Близькість порівнювальних таблиць визначається виразом:

$$TableP = (pMid_{row} + pMid_{column}) / 2 , \quad (6)$$

де $table_y \in Ttable_j$ та $table_y \in Ttable_{j+1}$

На основі близькості між таблицями двох порівнювальних документів T_j та T_{j+1} можна обрахувати їх загальну близькість на основі виразу:

$$proc_3 = (\sum_{p=1}^{k} TableP_p) / k , \quad (7)$$

де k – кількість порівнянь.

Встановлення зв'язків між наборами даних при створенні сховища наборів відкритих даних. Запропонований в роботі розвиток методу порівняння текстів шляхом визначення інтегральної близькості структурованих текстів та їх елементів у вигляді рядків і стовпців *CSV*-таблиці із ресурсів різних наборів відкритих даних було апробовано після включення його програмного забезпечення у вигляді додаткового модуля до програмної системи, представленої на рисунку 1.

Рис. 1 – Схема взаємодії програмних модулів автоматизованого створення сховища наборів відкритих даних

На веб-порталі Головного управління статистики в Одеській області, розміщеного за адресою http://od.ukrstat.gov.ua/, у розділі «Експрес-випуски» зберігаються *DOC*-документи з даними за різні роки та у різних категоріях, наприклад, освіта, ринок праці, зайнятість та безробіття, оплата праці та соціально-трудові відносини, соціальний захист, економічна діяльність, будівництво, внутрішня торгівля, діяльність підприємств, капітальні інвестиції, навколишнє середовище, послуги, наука, промисловість, сільське господарство, транспорт, зовнішньоекономічна діяльність. Їх *CSV*-еквіваленти створено вищезгаданими програмними модулями та розміщено на громадському веб-порталі відкритих даних Одеської області за адресою http://data.ngorg.od.ua.

Висновки: Проаналізовано 65 документів *DOC*-формату, які містили 143 таблиці з даними В результаті роботи модулів системи у сховищі даних було створено 143 набори відкритих даних у *CSV*-форматі. Для 87 наборів було автоматично встановлено семантичні зв`язки на рівні відкритих даних районів області. Знайдені набори створили сховище даних, яке дозволить проводити більш якісну аналітичну оцінку соціально-економічних процесів області з використанням діаграм та картографічних засобів візуалізації.

ВИКОРИСТАНІ ДЖЕРЕЛА

1. Методы и модели анализа данных: OLAP и Data Mining / [А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод]. – СПб. : БХВ-Петербург, 2004. – 336 с.

2. Постанова Кабінету Міністрів України «Про затвердження Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних» від 21.10.2015 № 835. : за станом на 1 грудня 2015 р. – Режим доступу : http://zakon3.rada.gov.ua/laws/show/835-2015-%D0%BF – Назва з екрана.

3. Скрипкін, М.О. Методика створення наборів відкритих даних на основі структурно-стилістичного аналізу електронних документів / М.О. Скрипкін, В.Г. Янко, А.А. Сіроцінський // Управління проектами, програмами, портфелями : Тези доповідей І Міжнародної науков-практичної конференції : [у 2т.]. – Том 1. – Одеса : Бондаренко М.О., 2016. – С.153 – 155.

4. Кунгурцев, А. Метод автоматизированного построения толкового словаря предметной области / А.Б. Кунгурцев, Я.В. Поточняк, Д.Ф. Силяев // Технологический аудит и резервы производства. 2015. – № 2/2(22). – С. 58 – 63.

5. Кунгурцев, Олексій. Побудова словника предметної області на основі автоматизованого аналізу текстів українською мовою / О. Кунгурцев, С. Ковальчук, Я. Поточняк, М. Широкоступ // Чернігів. Технічні науки та технології. – 2016. – № 3 (5). – С. 164 – 174.

Kungurtsev O.B., Blazhko O.A., Kovalchuk S.O., Skripkin M.O.

Establishment of Data Sets Relationships for Data Warehouse Creation

The paper considers the process of creating of electronic documents in data warehouse on the web-portal of open data.The result of this study is the method of text comparison for two structured electronic documents, which presented in tabular form to determine the possibility of their association in the data warehouse. The Scientific novelty of work is improvement of method for comparing the texts with integrated proximity of structured texts and their elements in rows and columns in a table, which allows to automate the process of establishing a semantic link between the data sets to create a data warehouse.