УДК 004.912

Бакай Д.М., Ярмолюк Р.С.

Хмельницький національний університет, Україна

РОЗБІР ОДНОРІВНЕВОГО БІБЛІОГРАФІЧНОГО ОПИСУ НА СКЛАДОВІ

Основною метою даного дослідження є розробка підходу до проектування підсистеми автоматичного розбору однорівневого бібліографічного опису у записах електронного каталогу бібліотеки за допомогою апарату регулярних виразів.

The main purpose of this study is to develop an approach to the design of the subsystem automatic analysis of single-level bibliographic records describing electronic library catalog using the apparatus of regular expressions.

Постановка проблеми

Робота з аналізу та пошуку помилок в записах електронного каталогу передбачає роботу з великими масивами текстових даних. Такі поля бібліографічного запису, як авторський знак, УДК та ББК індекси, ISBN - номер, мають певну визначену шаблонну структуру запису, перевірка якої не завжди інтегрована у систему перевірки коректності системи керування базою даних (СКБД) електронного каталогу. З іншого боку, поширеною проблемою у роботі з базою даних електронного каталогу є NULL-значення (тобто відсутність даних) у певних полях бібліографічного запису. Для вирішення даної проблеми відомий механізм запозичень інформації із зовнішніх джерел [1]. До таких джерел можна віднести:

- інформаційні ресурси мережі Інтернет;
- списки використаних джерел та переліки посилань у базах даних наукових статей;
- бібліографічні описи із зовнішніх баз даних видавництв та бібліотек.

До основних задач, що виникають при розробці систем верифікації даних електронного каталогу бібліотеки можна віднести [1]:

- Перевірка на коректність стандарту структурних (мають визначену структуру запису) атрибутів кортежу бази даних електронного каталогу. Індексні атрибути (авторський знак, УДК, ББК, ISBN, рік видання) мають наперед визначену структуру запису.

• Приведення даних до одного уніфікованого запису. Такі атрибути бібліографічного запису, як автор, назва, видавництво, тощо мають довільну структуру запису, але для ефективного пошуку та обробки даних необхідно привести всі дані одного атрибуту до певного шаблону запису. Зазвичай засобів самої СКБД недостатньо.

• Розбір бібліографічного опису на складові, або у загальному, виокремлення значень із атрибутів вільного формату (розщеплення атрибутів). Сам електронний каталог містить у собі велику кількість необробленої та неструктурованої інформації. Зокрема, списки використаних джерел та переліки посилань, що містяться у наукових статтях, монографіях, навчальних посібниках. Структура таких записів визначена відповідними нормативними документами. Тому пошук таких структур у тексті за певним шаблоном та виокремлення необхідних атрибутів для запису у базу даних є актуальною проблемою.

• Запозичення бібліографічної інформації з мережі Інтернет. Задачу доповнення відсутніх даних можливо вирішити за допомогою інформаційного поля мережі Інтернет. На основі відомої інформації формується пошуковий запит для певної пошукової машини (Google, Яндекс, Mail.ru, тощо). Отриманий результат являє собою документ у розмітці HTML.

Зазвичай однорівневий бібліографічний опис (надалі бібліографічний опис) який знаходиться в базі даних зберігається в незручному для автоматичної обробки вигляді. Згідно [2] (ДСТУ ГОСТ 7.1: 2006 "Система стандартів з інформації, бібліотечної та видавничої справи. Бібліографічний запис. Бібліографічний опис. Загальні вимоги та правила складання") він повинен представлятись у вигляді такої структури (рис. 1).

У структурі (рис.1), деякі елементи є необов’язковими, а деяких зовсім може не бути, проте є основні, яких в тій чи іншій мірі достатньо для ідентифікації певного документа [3]. І тут постає ряд питань, які потребують відповіді. Наскільки ця інформація придатна для автоматичної обробки? Чи можна і чи потрібно її застосовувати?

Формулювання цілей та мети дослідження.

Основною метою даного дослідження є розробка ефективних методів та засобів розв’язання проблеми проектування підсистеми автоматичного розбору однорівневого бібліографічного опису у записах електронного каталогу бібліотеки за допомогою апарату регулярних виразів.

Виклад основних матеріалів дослідження

Бібліографічний опис в найбільш широкому спектрі складається з восьми частин (зон), однак в більшості випадків їх менше як вісім, це пов'язано з тим, що деякі частини (зони) є факультативними (необов'язковими).

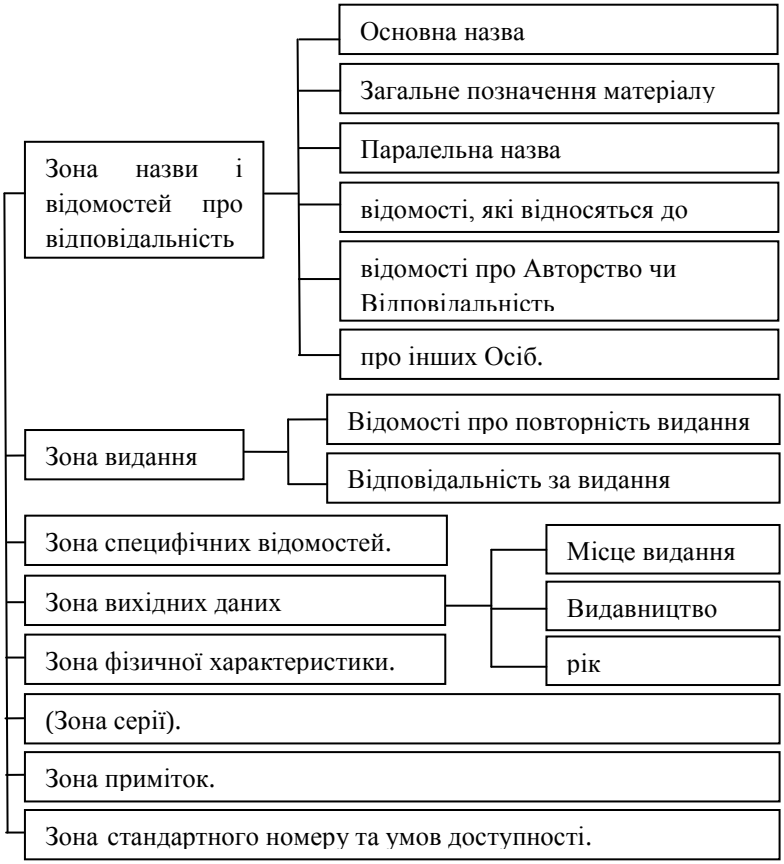

Рисунок 1 – Структура однорівневого бібліографічного опису.

В свою чергу всі ці зони поділяються на під-зони (елементи), наприклад зона вихідних даних складається з місця видання, видавництва та року видання (рис.1).

Перед елементами та зонами у бібліографічному описі ставлять знаки приписаної пунктуації. На відміну від звичайних

граматичних знаків, знаки приписаної пунктуації виконують розпізнавальні функції зон та елементів.

Загальний принцип механізму розбору бібліографічного опису.

Розбір бібліографічного опису буде складатись з трьох етапів:

- Попередня обробка;
- Лексико-синтаксичний аналіз;
- Пост-обробка.

Попередня обробка. Перед тим як приступити до розбору бібліографічного опису на складові насамперед потрібно привести його до запису одною стрічкою, тобто символи переходу на нову стрічку та повернення каретки, замінити пустим символом (пропуском).

Лексико-синтаксичний аналіз. Насамперед потрібно рядок бібліографічного опису розкласти на зони, на перший погляд це не є складною задачею оскільки кожна зона між собою розділена крапкою й тире, практично це можна зробити за допомогою більшості мов програмування в результаті ми одержимо масив записів. Складність настає тоді коли нам необхідно визначити, які записи до якої зони віднести, і яких нема, якщо такі є. Можна запропонувати наступну послідовність дій:

1. Перший запис у нашому масиві завжди буде відноситись до зони назви і відомостей про відповідальність, так як тут міститься назва документу, який і являється об'єктом опису;
2. Другий запис нашого масиву будемо відносити до зони видавництва, якщо у його контексті будуть зустрічатись такі слова як: «видавництво», «версія», «варіант» і т. п., а також їх скорочення та еквіваленти на інших мовах.
3. Зону специфічних відомостей важко описати, тому будемо шукати запис, який належить зоні вихідних даних яка характеризується таким шаблоном «Місце видання : Вид-во, рік», а на мові регулярних виразів [4] це виглядає так: **Regex(@"^.+?\s\:\s.+?\,\s[\[\w\.\s]*[\dMCLXIVD]+.*"),** знайшовши необхідний запис, та його порядковий номер у масиві, необхідно проаналізувати записи у масиві, які містяться між ним і першим записом, якщо таких записів немає то зон видавництва та специфічних відомостей не має, якщо є один запис і він підходить по перевірці другого пункту цього алгоритму то цей запис відноситься до зони видавництва, інакше до зони специфічних відомостей, якщо є два записи, то перший з них належить зоні видавництва, а наступний зоні специфічних відомостей;

4. До зони фізичної характеристики відносять наступний запис, який знаходиться після зони вихідних даних, якщо він починається з необов'язкової квадратної дужки що відкривається, за якою йдуть арабські або римські цифри;

5. До зони серії належить той запис, який взятий у круглі дужки;

6. Якщо запис починається з ISBN, ISSN, номеру державної реєстрації, видавничого номеру, інших номерів, а також слід враховувати той факт коли запис знаходиться в кінці масиву записів то цей запис відноситься до зони стандартного номеру та умов доступності;

7. Записи, які залишились будемо відносити до зони приміток.

Надалі, деякі зони можна розкласти на підзони, використовуючи шаблони регулярних виразів.

Розкладемо зону назви та відомостей про відповідальність на складові.

Наприклад зона назви та відомостей про відповідальність має такий запис: «Обычаи поволжских немцев [Текст] = Sitten und Brauche der Wolgadeutchen / Екатерина Ерина, Валерия Салькова ; худож. Н. Стариков ; [Междунар. союз нем. культуры].».

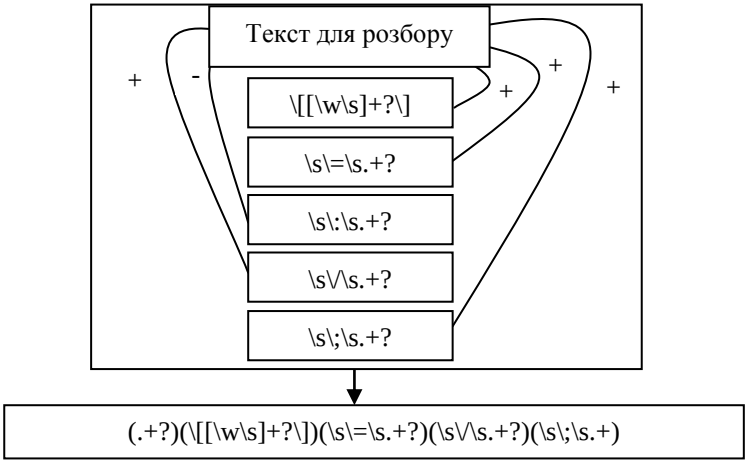

Рисунок 2 – Формування регулярного виразу.

Створюємо регулярний вираз, який описує всю зону *(.+?)*, та декілька шаблонів, які провіряють чи є в даній зоні відповідні знаки

приписаної пунктуації, якщо шаблон регулярного виразу повертає позитивний результат то цей шаблон додається до загального регулярного виразу. Отримаємо регулярний вираз, який описує 5 елементів нашої зони (рис.2).

В результаті застосування відповідного шаблону регулярного виразу та виконання пост-обробки, одержимо наступне (рис.3).

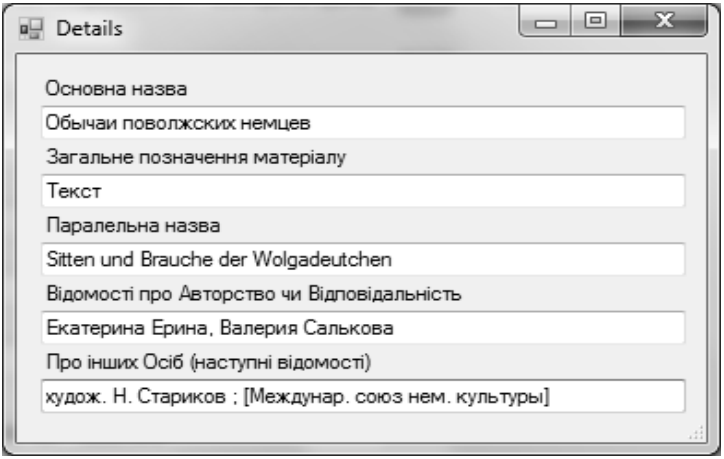

Рисунок 3 – Розбір зони назви та відомостей про відповідальність на елементи.

Пост-обробка. Отримані нами записи можуть містити на початку та кінці рядка такі символи: (' ', '/', '=', ':', ';', ',', '[', '.',']'), які необхідно видалити для коректності результату.

Висновки. Отже, за наведеним вище алгоритмом дій, бібліографічний опис можна автоматично розбирати, застосовуючи для цього, наприклад, апарат регулярних виразів, який підтримується більшістю мов програмування.

Надалі можна перевіряти коректність атрибутів з наперед визначеною структурою, до таких атрибутів належать ISBN, ISSN та ін. дані, які ми отримали з бібліографічного опису.

Література

1. Ярмолюк Р.С. Уніфікація атрибутів кортежу бази даних засобами регулярних виразів на прикладі електронного каталогу бібліотеки / Р.С. Ярмолюк //Вісник Хмельницького Національного Університету, серія Технічні науки, - 2012. - № 1 – С.186-189.

2. ДСТУ ГОСТ 7.1:2006. Бібліографічний запис, бібліографічний опис. Загальні вимоги та правила складання : метод. рекомендації з впровадження / уклали: Галевич О. К., Штогрин І. М. – Львів, 2008. – 20 с.

3. Вершинин М.И. Электронный каталог проблемы и решения / Вершинин М.И. – СПб. :ПРОФЕССИЯ, 2007. – 233 с.

4. Фридл Дж. Регулярные выражения, 3_е издание. – Пер. с англ. – СПб.: Символ_Плюс, 2008. – 608 с., ил.