ЗАСТОСУВАННЯ АПАРАТУ РЕГУЛЯРНИХ ВИРАЗІВ ПРИ РОЗРОБЦІ СИСТЕМИ ВЕРИФІКАЦІЇ ДАНИХ ЕЛЕКТРОННОГО КАТАЛОГУ

Ярмолюк Р. С.
Хмельницький національний університет

Від якості даних, що містяться в електронному каталозі напряму залежить якість інформаційно-пошукових послуг бібліотеки. За час свого життєвого циклу у електронному каталозі відбуваються певні інформаційно-технологічні процеси, що на пряму впливають на виникнення помилок в записах бази даних. Однак, аналіз програмного забезпечення для роботи з електронним каталогом у автоматизованих бібліотечних інформаційних системах (АБІС) показав відсутність функцій для верифікації та уточнення даних. Тому, розробка системи контролю за якістю даних електронного каталогу бібліотеки є важливою технічною проблемою.

Проблеми пошуку та корекції орфографічних помилок в записах електронного каталогу у своїх працях підіймати такі вчені, як Вершинин М.И., Крауш А.С., Nielsen R., Ballard T., Randall B. та інші [1-5]. У своїх працях та розробках дані автори приділяли увагу розробці методів та засобів пошуку та корекції орфографічних помилок у бібліографічних записах. Але питання структурної верифікації полів запису та розробки механізмів доповнення та уточнення відсутніх даних залишаються відкритими.

Метою даної роботи є розробка основних принципів та методик застосування апарату регулярних виразів при проектуванні системи верифікації даних електронного каталогу (СВДЕК).

Виклад основного матеріалу

Система верифікації даних електронного каталогу – набір програмних засобів для пошуку, оцінки, виправлення та уточнення помилкових даних в електронних каталогах бібліотек [6]. Структурно-функціональна схема СВДЕК представлена на рисунку 1.

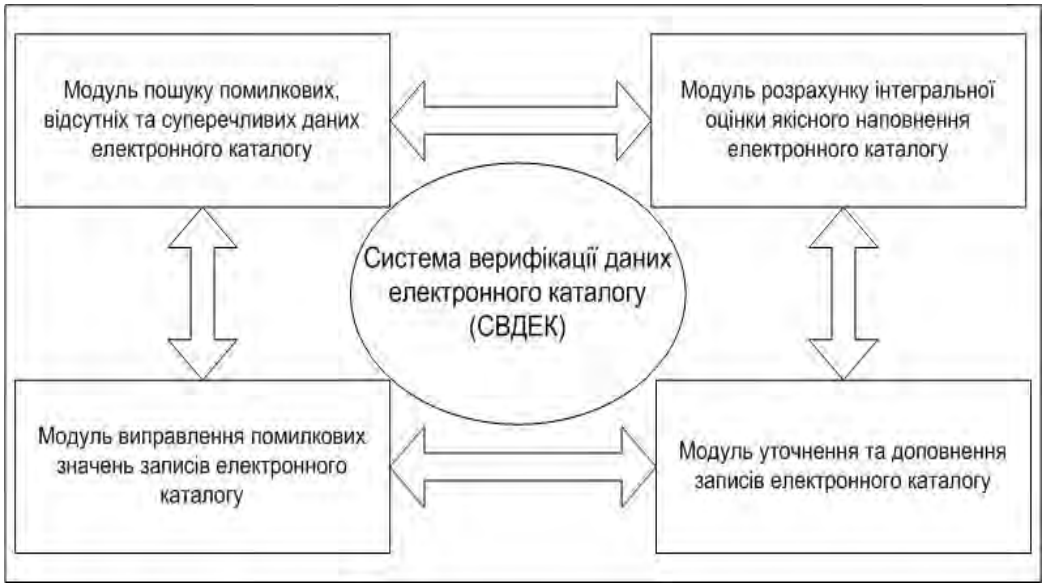

Рис.1 Структурно-функціональна схема СВДЕК.

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 запису 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 запропоновано механізм запозичень інформації із зовнішніх джерел. 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.

Якщо для імпорту даних із зовнішніх електронних каталогів існують інструментарії обробки у структурі АБІС, то для мережі Інтернет та списків літератури необхідно застосовувати підходи синтаксичного аналізу. Зокрема, розробка модулів-парсерів HTML сторінок є досить трудомісткою роботою.

Для викладених вище задач при розробці відповідних модулів СВДЕК пропонується використовувати математичний апарат регулярних виразів та теорію формальних мов.

Регулярний вираз – це формальна мова пошуку і проведення маніпуляцій з підрядками в тексті, заснована на використанні метасимволів, або у контексті механізму шаблонування, - це рядок, що описує або збігається з множиною рядків, відповідно до набору спеціальних синтаксичних правил [7].

По суті, регулярні вирази – це простий та зручний спосіб запису регулярних множин. у вигляді звичайного рядка. Більш докладно про математичні основи та теорію формальних мов можна ознайомитись у [8]. З точки зору застосування регулярний вираз задає зразок пошуку. Після чого можна перевірити, чи задовольняє заданий рядок або його підрядок даному зразку.

Синтаксис регулярних виразів залежить від платформи або мови програмування на якій реалізується програмне забезпечення та застосовує різний набір засобів, зокрема[7]:

- символи і escape-послідовності;
- символи операції і символи, що позначають спеціальні класи множин;
- імена груп і обернені посилання;
- символи тверджень та інші засоби.

На даний час переважна більшість програмних платформ та мов програмування реалізують у своїх можливостях підтримку регулярних виразів. Серед них Perl, Java, PHP, JavaScript, мови платформи .NET Framework, Python, Ruby, та інші. Наприклад для об’єктно-орієнтованої мови програмування C# дані можливості реалізовані у просторі імен ***System.Text.RegularExpressions.***

З переліком символів та правилами побудови регулярних виразів для різних платформ та мов програмування представлено у [7].

Отже, визначимо задачі, які ефективно вирішуються допомогою регулярних виразів:

- Перевірка на коректність шаблону структурних (мають визначену структуру запису) атрибутів кортежу бази даних електронного каталогу. Зокрема, для індексних атрибутів (авторський знак, УДК, ББК, ISBN, рік видання) за допомогою регулярних виразів реалізується перевірка на коректність.
- Приведення даних до одного шаблону запису. Зокрема, 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, тому

реалізація шаблонування та стандартизації ефективно проводиться за допомогою регулярних виразів.

- Розбір бібліографічного опису на складові. 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, тощо. 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 розв'язною проблемою за допомогою апарату регулярних виразів.
- Запозичення бібліографічної інформації з мережі Інтернет. Задачу доповнення відсутніх даних можливо вирішити за допомогою інформаційного поля мережі Інтернет. Зокрема на основі відомої інформації формується пошуковий запит для певної пошукової машини (Google, Яндекс, Mail.ru, тощо). Отриманий результат являє собою документ у розмітці HTML. Далі за допомогою регулярних виразів проводиться синтаксичний аналіз та відокремлюються дані необхідні для інтеграції.

Запропонована система верифікації даних електронного каталогу, що повній мірі забезпечує усі необхідні функціональні можливості для підтримки якості даних у електронному каталозі. У основу концепції розробки модуля уточнення та доповнення записів каталогу покладено апарат регулярних виразів. Визначено основні проблеми, що виникають при роботі даного модуля, та запропоновано шляхи їх розв'язання за допомогою апарату регулярних виразів. Подальші дослідження будуть спрямовані на побудову основних регулярних виразів для перевірки на коректність шаблону, приведення даних, реалізацію механізмів запозиченні відсутніх даних із різних джерел та дослідження ефективності даних методів на реальних даних.

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

2. Крауш А. С. Утилиты для проверки и коррекции электронных каталогов / А. С. Крауш, Д. Ю. Копытков, А. С. Макаревич // Библиотечное дело. – 2005. – № 6 (30). – С. 21 – 24.

3. Nielsen R. Lost articles: Filing problems with initial articles in data bases / R. Nielsen, J. M. Pyle // Libr. Resources a. techn. Services. – 1995. – Vol. 39, №3. – P. 291 – 293.

4. Ballard T. Spelling and typographical errors in library databases: one libr. System for noting out spelling error / T. Ballard // Computer in libr. – 1992 – Vol. 12, №6. – P.14 – 19.

5. Randall B. Spelling Errors in the Database: Shadow or Substance? / B. Randall // Libr. Resources a. techn. Services. – 1999. – Vol. 43, №3. – P. 161– 170.

6. Ярмолюк Р.С. Структурно-функціональна модель системи верифікації даних електронного каталогу / Р.С. Ярмолюк // Сучасні проблеми діяльності бібліотеки в умовах інформаційного суспільства: матеріали третьої науково-практичної - Львів: 2011. – С. 217-224.

7. .Friedl J. Mastering Regular Expressions Third Edition / J. Friedl. – California: O'Reilly Media, 2006. – 534p.

8. Пентус А.Е. Теория формальных языков. Учебное пособие / А.Е. Пентус, М.Р. Пентус. – Москва: МГУ, 2004. – 80с.