УДК 004.62::01

МЕТОД ВИЯВЛЕННЯ ПОДІБНОСТІ НАЗВ ВИДАВНИЦТВ, СТІЙКИЙ ДО РІЗНИХ ВИДІВ СКОРОЧЕНЬ

DOI 10.36994/2788-5518-2021-02-02-16

Таланчук П.М., *д.т.н., проф. Відкритий міжнародний університет розвитку людини «Україна»*

Петренко М.В.*, Відкритий міжнародний університет розвитку людини «Україна», mxms@ukr.net*

Анотація. *Стаття присвячена вдосконаленню методів інтеграції гетерогенних бібліографічних даних на основі скоординованого використання полів «ISBN» та «Видавництво». Вдосконалення стосується частини порівняння назв видавництв та має на меті усунути виявлені на реальних даних слабкі місця.*

Інтеграція бібліографічних даних з різних джерел використовується з метою отримання узгоджених бібліографічних ресурсів на їх основі та для підвищення загальної якості даних. Такі процеси актуальні як в Україні, так і в світі. Сьогодні в Україні є запит на цифрову трансформації бібліотек та створення зведених електронних каталогів.

В ході дослідження розглянуті методи інтеграції у вигляді, в якому вони зараз застосовуються в передових світових дослідженнях. Проаналізована можливість їх застосування до бібліографічних даних публічних бібліотек м. Києва (141-а бібліотека) та проблеми, які можуть виникнути.

В ході розгляду було помічено, що в значеннях поля «Видавництво» у великій кількості присутні різні види скорочень. Цей факт заважає ефективному застосуванню методів нечіткого порівняння текстових даних, які використовуються в методах інтеграції бібліографічних даних (відстань редагування Левенштейна, метод Яро-Вінклера, метод 3-грам).

В результаті аналізу бібліографічних даних було розроблено рекурентний метод визначення того, що два рядки можуть бути варіантами написання назви одного видавництва, з врахуванням можливості використання скорочень. Метод описано у вигляді блок-схем.

На основі пропонованого методу та згаданих вище методів нечіткого порівняння текстових даних розроблено програмне забезпечення, яке протестоване на даних публічних бібліотек Києва. Доведено, що пропонований метод розширює можливості розпізнавання варіантів назв видавництв чим може підвищити ефективність інтеграції даних. Запропоновані шляхи використання авторського методу разом з іншими, а також напрямки його розвитку.

Ключові слова: *бібліографічні дані, визначення подібності текстових даних, скорочення, абревіатура, назва видавництва.*

THE PUBLISHER NAMES MATCHING METHOD, RESISTANT TO DIFFERENT KINDS OF ABBREVIATIONS

***Petro Talanchuk**, Dr. habil., Prof. Open International University of Human Development "Ukraine"*

***Maxim Petrenko**, Open International University of Human Development "Ukraine", mxms@ukr.net*

***Abstract.** The article is devoted to the improvement of methods for integrating heterogeneous bibliographic data based on the coordinated use of the fields "ISBN" and "Publisher". The improvement concerns part of the comparison of publishers names and is aimed at eliminating weaknesses identified on real data.*

The integration of bibliographic data from different sources is used to obtain consistent bibliographic resources from them and to improve the overall quality of the data. Such processes are relevant both in Ukraine and in the world. Today in Ukraine there is a request for the digital transformation of libraries and the creation of consolidated electronic catalogs.

In the course of the study, the methods of integration are considered in the form in which they are now used in advanced world researches. The possibility of their application to the bibliographic data of public libraries in Kiev (141 libraries) and possible problems are analyzed.

In the course of the proceedings, it was noticed that in the values of the field "Publisher" there are a large number of different types of abbreviations. This fact interferes with the effective application of methods of fuzzy text data comparison used in integration methods (Levenshtein editing distance, Jaro-Winkler method, 3-gram method).

As a result of the analysis of bibliographic data, a recursive method was developed to determine that two lines can be spellings of the name of one publisher, taking into account the possibility of using abbreviations. The method is described in the form of block diagrams.

Based on the proposed method and the above-mentioned methods of fuzzy comparison of text data, software has been developed that has been tested on the data of public libraries in Kyiv. It is proved that the proposed method expands the possibilities of recognizing variants of publishers names, which can increase the efficiency of data integration. The ways of using the method together with others and directions of its development are proposed.

***Keywords:** bibliographic data, text data similarity, abbreviation, publisher name, fuzzy comparison.*

Вступ

Інтеграція бібліографічних даних отриманих з різних джерел необхідна для створення якісних інформаційних ресурсів, які мають підтримувати діяльність бібліотек та допомагати забезпечувати інформаційну рівність громадян. Бібліографічним даним характерна різнорідність, яка створюється не лише при об'єднанні даних з різних джерел, а і на етапах виготовлення макету та реєстрації книги в бібліографічній базі.

Найважливішими проблемами, які стоять перед вдосконаленням інформаційних бібліографічних ресурсів (особливо – зведених) є видалення

дублів та виправлення помилок в даних, для підвищення однорідності та зменшення надмірності.[1]

Такі властивості книги, як назва, автор, код класифікатора мають стосунок до твору. Різні видання одного твору прийнято розділяти через те, що вони можуть володіти відмінними споживчими характеристиками. До видання мають стосунок: назва видавництва, ISBN, рік видання, кількість сторінок.

Для розпізнавання бібліографічних записів, які стосуються одного і того ж видання, важливим є визначення видавництва (видавництв), до якого вони належать. Дані про видавництво вводяться в різній формі, і не завжди між ними є відповідність. Різниця може виникати як на етапі підготовки книги, так і на етапі введення даних в бібліографічну базу даних.

Для досягнення мети інтеграції бібліографічних даних необхідно встановити відповідність між різними варіантами назв одного і того ж видавництва. Додатковою користю від цього може стати можливість організовувати чи фільтрувати книги за видавництвами в бібліотечних інформаційних ресурсах без впливу проблем викликаних низькою якістю даних.

Існує кілька основних джерел різниці між написанням назви одного і того ж видавництва:

- видавництво було перейменовано;
- назва написана іншою мовою;
- наявні помилки введення;
- наявні скорочення слів назви.

Після перейменування назви можуть суттєво відрізнятися, хоча для одного і того ж видання це не характерно. Більшість назв написаних іншою мовою для порівняння потребують лише транслітерації і більшої толерантності до одиночних символьних розбіжностей.

У бібліографічних базах даних м. Києва найбільш поширеними варіаціями назв видавництв є: назви введені з помилками, назви введені з скороченнями.

Огляд літератури

Дослідники міжнародної науково-дослідної організації OCLC (Online computer Library Center) працювали над розробкою методів ідентифікації різних варіантів назв видавництв та створення авторитетного файлу видавництв – PNAF (Publisher Name Authority File)[2], застосовуючи для групування видавців відповідний префікс ISBN. Для оцінювання подібності назв застосовувався метод 3-грам та визначення відстані редагування Левенштейна.

Зараз дослідження OCLC відтворює та продовжує Джим Хахн – голова «Metadata research» Університету Пенсильванії. В своїй роботі він використовує бібліотеку OpenRefine, в якій застосовуються ті ж самі методи визначення подібності двох рядків – n-грами та відстань редагування Левенштейна.[3]

Такі методи є загальновживаними для порівняння рядків [4, 5] і конкретно для порівняння назв компаній [6, 7]. Згадані методи порівняння добре працюють при невеликих різницях в текстах, що порівнюються, наприклад, при наявності одиночних помилок введення.

В результаті дослідження реальних даних поля «Видавництво» публічних бібліотек м. Києва, було виявлено, що серед бібліографічних даних, що порівнюються, наявно достатньо багато різних варіантів скорочень, які мають значно більшу символьну різницю та вимагають застосування інших методів порівняння.[8]

Відомі методи отримання сталих скорочень [9, 10] з:

- пояснень розміщених в тексті в дужках : «OCLC (Online Computer Library Center)»;
- із зносок;
- інших джерел, де ідентичність чітко вказана.

Відмінність проблеми наявності скорочень у застосуванні до порівняння назв видавництв:

сталі скорочення найчастіше є малозначимими частинами назви («ФОП», «ООО», «ТОВ» і т.д.). Відповідно, використання попередньо отриманих сталих скорочень не вирішує проблеми. Рішенням можуть стати методи, які не використовують базу скорочень, а використовують правила скорочення;

необхідність не відновлювати слова з скорочень, а лише встановити можливість того, що два рядки можуть бути варіантами написання однієї назви з можливим застосуванням скорочень.

Подібний метод (за умови достатньої швидкодії) може бути застосований не лише на етапі обробки введених бібліографічних даних, а і на етапі введення в систему (надання пропозицій варіантів назв, які можуть відповідати назві видавництва, яка вводиться користувачем), щоб запобігати помилкам введення. На етапі введення користувачу буде не важко відповісти ствердно в разі наявності реально збігу та допомогти цим формуванню якісного авторитетного файлу назв видавництв.

Виконання досліджень

Робота над створенням методу порівняння назв проводилася на основі реальних даних та з врахуванням логіки утворення скорочених варіантів назв. В реальних бібліографічних даних крапки недостатньо часто ставлять в кінці слова, що було скорочене. Також великі літери не завжди позначають перші літери слів при утворенні скорочення з кількох слів. Натомість крапка може бути наявна після слів, які не було скорочені, а великі літери в словах, які не є скороченнями. У зв'язку з цим покладатися на такі ознаки не можна. А отже, кожне слово потенційно може бути скороченням, незалежно від ознак.

Зустрічаються як звичайні ініціальні абревіатури типу:

- «МАУП»;
- «Міжрегіональна Академія Управління Персоналом».

- Складені:
- «Держвидав худліт»;
- «Державне видавництво художньої літератури».
- Скорочення з дефісом:
- «Держвидав худож. літ-ри»;
- «Держ. видавництво художньої літератури».

В таких скороченнях можна вважати, що слово закінчується на дефісі, оскільки дефіси відсутні у абревіатурі, а отже літери після дефіса є закінченням слова і не мають використовуватися для подальшого порівняння.

Важлива кожна літера скорочених слів:

- держ – держвидав;
- худ – художньої;
- літ. – літератури;

У випадку скорочень з дефісом, важливі лише літери до дефіса. Слово з яким відбувається порівняння може бути скороченим чи повним. Однакове закінчення не може надійно підтвердити, що знайдено повну форму слова. Крім того, закінчення є більш варіативними частинами слова, ніж його початок. На пошук неповної відповідності, яка нам потрібна, закінчення не впливає: літ-ри - літератури

Для назв видавництв характерна стала послідовність слів. Також перші літери слів вводяться більш ретельно, ніж наступні. слів і Це можна використати у проектуванні методу.

Основний вид помилок, які характерні для перших літер – введення візуально подібного символу з використанням іншої розкладки клавіатури. Такі випадки легко можна виявляти, а помилки автоматично виправляти (в АБІС таких засобів немає, але в роботі[11] описано простий для обчислень метод). Тому будемо вважати, що помилки такого виду виправлені до проведення порівняння назв.

Метод має змогу перевіряти можливість збігу не лише між скороченим та повним варіантами написання, але і між різними варіантами скорочень, що відповідають одному слову (останній варіант при кількісних методах порівняння має найменший відсоток збігу).

Метод описано у вигляді взаємно пов’язаних рекурентних функцій «Порівняння обох слів», «Пошук по Н1» (назві 1), «Пошук по Н2» (назві 2), блок-схеми яких знаходяться на рисунках 2 - 4.

Завдання методу – відповісти на питання, чи можливо, що порівнювані рядки є варіантами одного тексту з врахуванням можливості використання в них різних варіантів скорочень. Старт спроектованого методу порівняння показано на рис. 1.

Скорочення тексту суттєво зменшує кількість інформації і відповідно збільшує невизначеність (особливо при утворенні ініціальних скорочень). Одна абревіатура може за всіма ознаками бути варіантом написання різних назв. Вона навіть може в різних випадках представляти одну назву, а в інших

– іншу. Без врахування додаткових параметрів не лише алгоритм, але і спеціаліст-бібліограф у значній кількості випадків не зможе відповісти точніше ніж варіант можливий або ні.

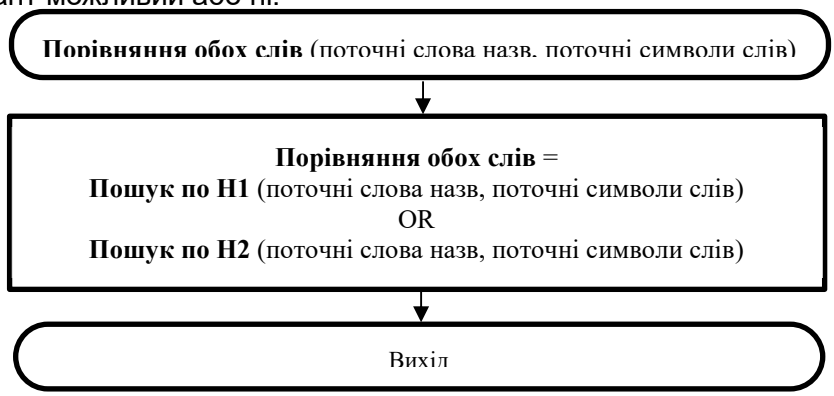

Рис. 1: Блок-схема роботи запуску пропонованого методу порівняння назв.

Особливість порівняння значень бібліографічного поля «Видавництво» в тому, що у бібліографічній базі даних є достатньо багато кодів ISBN. Якщо відновити їх структуру, то стає можливо згрупувати бібліографічні записи за префіксом, який закінчується блоком видавництва.[12, 2] Таким чином можна порівнювати між собою лише значення полів, які належать до одного блоку видавництва і таким чином отримати додаткову інформації, якої бракує у зв'язку з втратою літер при скороченні.

Оцінка ефективності запропонованого методу

Описаний вище метод був реалізований на мові Python та застосований до бібліографічних даних отриманих з електронних каталогів [13, 14]. Дані були очищені від візуально непомітних помилок введення, та уніфіковані. В процесі порівняння дані очищувалися від неалфавітних символів та слова з дефісами скорочувалися до дефісу.

Усього записів видавництв в момент проведення тестування – 420215. З них унікальних – 22402 записи. Записи попередньо відсортовані за алфавітом. Перевірка проводиться в межах збігу першої літери назв.

Проведено близько 20 млн попарних порівнянь. Після завершення порівнянь видалено симетричні пари (метод порівняння є симетричним і перестановка порівнюваних значень не впливає на результат), залишилося 6518 унікальних пар, які можуть бути варіантами написання однієї назви.

Для всіх заданих пар проведено ручну перевірку точності розпізнавання та виставлені наступні варіанти нечіткої відповідності:

точно стосується одного видавництва (3539 – 54%);

дуже ймовірно, що стосується одного видавництва (528 – 8%);

можливо, стосується одного видавництва, але з неточністю чи помилкою (1038 – 16%);

дуже ймовірно, що це назви різних видавництв (1413 – 22%).

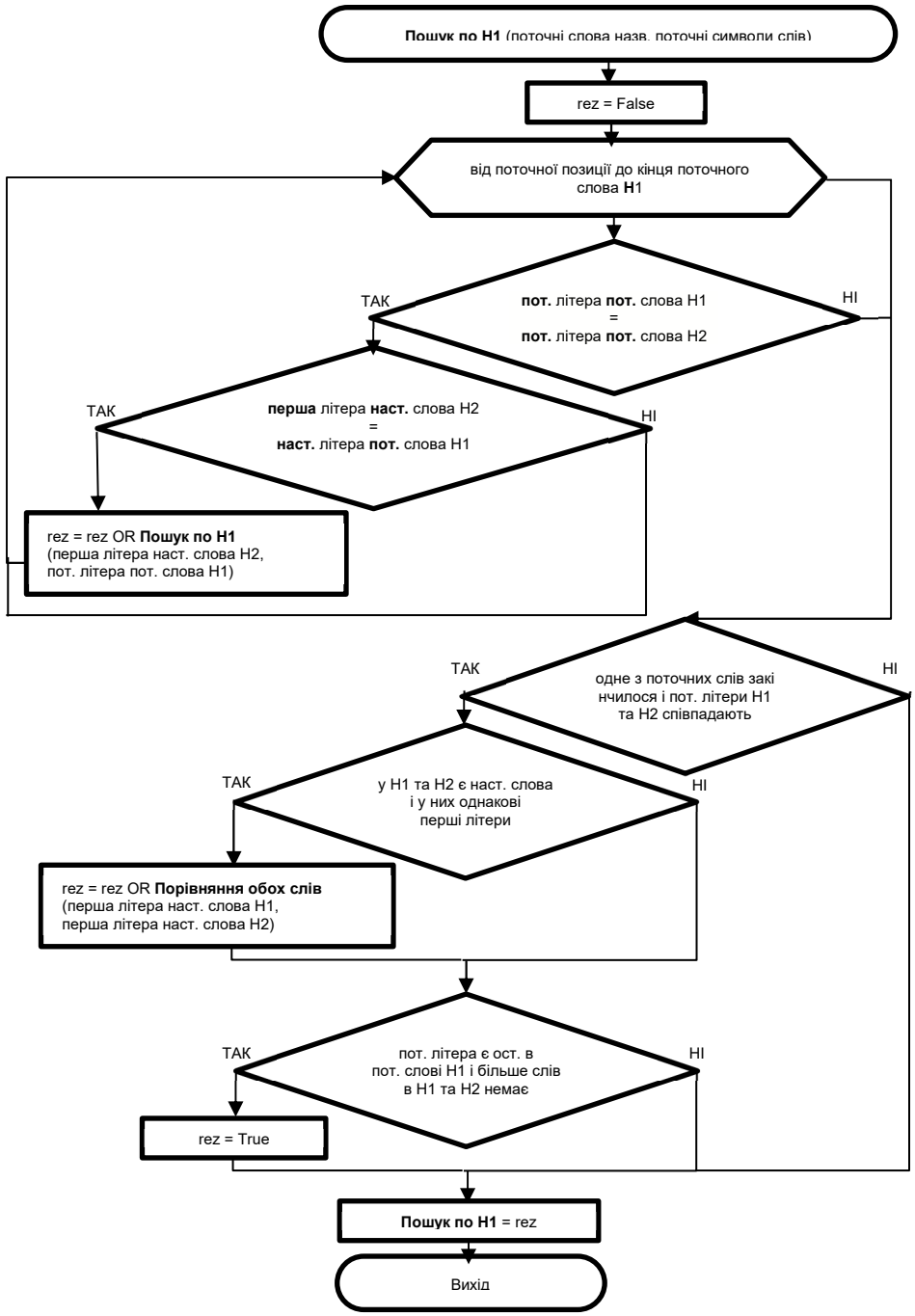

Рис. 2: Блок-схема роботи пропонованого методу порівняння («Пошук по Н1»).

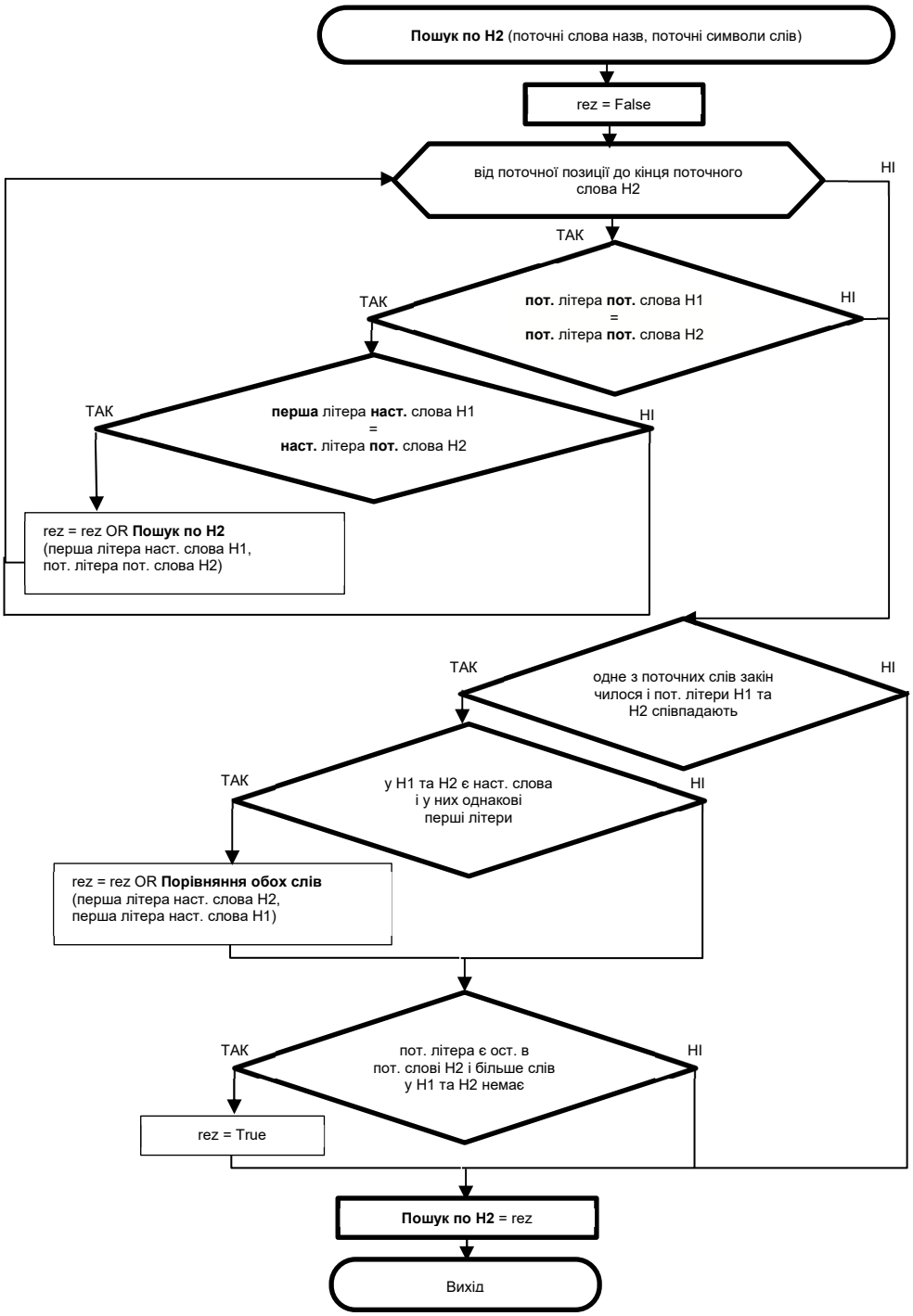

Рис. 3: Блок-схема роботи пропонованого методу порівняння («Пошук по Н2»).

Перші 2 варіанти можна вважати вірно-позитивним (True Positive) результатом, третій – невизначеним (ND), а четвертий – хибно-позитивним (False Positive). Перевірити всі сотні тисяч порівнюваних пар, які не пройшли порівняння та визначити чи є серед них хибно-негативні пари не було можливим. Було досліджено відкинуті пропонованим методом пари з найвищою мірою подібності за методом Яро-Вінклера (90% подібності та більше). У цій множині були виявлені лише пари, які відрізняються помилками чи додатковими словами. Пропущених варіантів скорочень без помилок не виявлено.

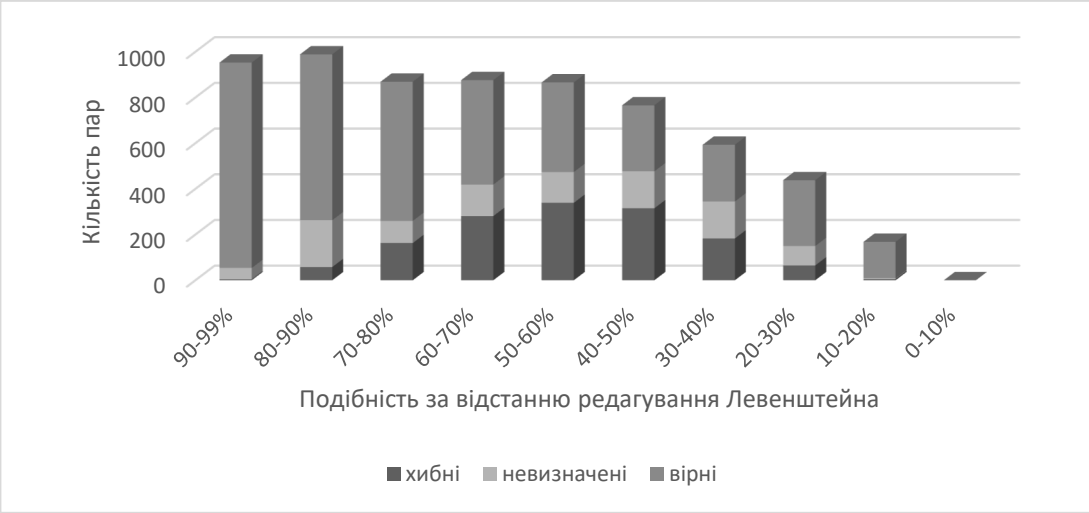

Рис. 4: Розподіл вірно, хибно та недостатньо точно визначених пар за відстанню редагування Левенштейна.

Рис. 5: Розподіл вірно, хибно та недостатньо точно визначених пар за методом Яро-Вінклера.

Паралельно до визначення подібності пар розробленим методом, вимірювалася міра подібності за нечіткими методами порівняння Яро-Вінклера, 3-грам та відстані редагування Левенштейна.

На рисунках 4-6 розглянуто пари, що пройшли перевірку пропонованим методом в розрізі відсотків подібності за відповідними методами.

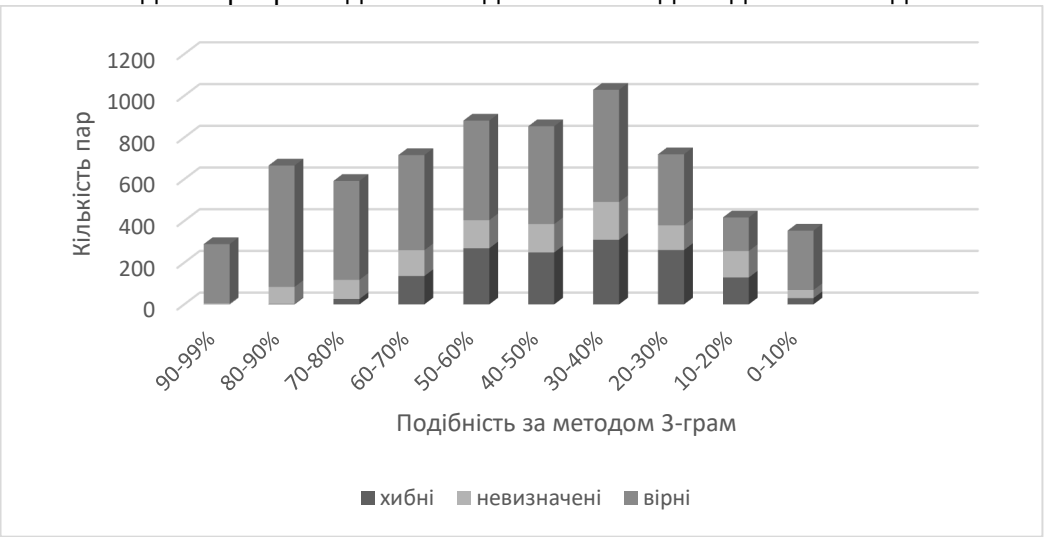

Рис. 6: Розподіл вірно, хибно та недостатньо точно визначених пар за методом 3-грам.

В розрізі усіх нечітких порівнянь помітно, що за високого та низького рівня подібності ми маємо найменше хибно-позитивних спрацювань. Приклад пари з подібністю 16% за відстанню редагування:

- «МЕЖРЕГІОНАЛЬНА АКАДЕМІЯ УПРАВЛІННЯ ПЕРСОНАЛОМ»;
- «МАУП».

Приклади пар з подібністю 40-43%:

- «ГОЛОВНА РЕДАКЦІЯ УРЕ»;
- «ГОЛ. РЕД. УКРАЇНСЬКОЇ РАДЯНСЬКОЇ ЕНЦИКЛОПЕДІЇ»;

та

- «ГОСУДАРСТВЕННОЕ ИЗДАТЕЛЬСТО ХУДОЖЕСТВЕННОЙ ЛИТЕРАТУРЫ»;
- «ГОСИЗДАТ. ХУД. ЛИТ».

Деякі з знайдених пар лише фахівець може визначити як різні – «ДІЛОВИЙ ПАРТНЕР» та «ДІПА». Або ж такі, які неможливо однозначно спростувати чи підтвердити без додаткової інформації – «PROVIDENCE ASSOCIATION» та «PA».

Метод порівняння Яро-Вінклера більше враховує правильність перших літер слова, ніж наступних. Такий підхід актуальний не лише у зв'язку з тим, що помилки в перших літерах слів трапляються рідше, а і тому, що різні види скорочень також зберігають перші літери слів.

Проте багато встановлених пропонованим методом пар мають дуже низьку оцінку подібності Яро-Вінклера, як і розраховану на основі інших методів. Подібність «Видавництво Старого Лева» та «ВСЛ» – 46%, а «Львівській національній університет імені Івана Франка» та «ЛНУ ім. І. Франка» - 67% подібності.

Отже, позбавитися від помилково-позитивних спрацювань метод Яро-Вінклера допомогти не може. Подібність «Діловій партнер» та «ДІПА» за методом Яро-Вінклера - 80%, що вище, ніж подібність за відстанню редагування Левенштейна (42%). Цей метод не може ні замінити пропонований у даному застосуванні, ні допомогти уникнути помилково-позитивних спрацювань. Найкраще застосування методу Яро-Вінклера – співставлення назв з звичайними скороченнями:

- «Государственное издательство Художественная литература»
- «Гос. изд. худож. літ.»

Методи нечіткого порівняння Левенштейна, Яро-Вінклера та 3-грам мають високу точність порівняння лише на високих (90-95% відповідно) мірах подібності. При зниженні відсотку понад 80-90% відповідно, кількість хибно розпізнаних пар стає занадто великою для їх використання.

Про корисність запропонованого методу свідчить те, що 81-85% розпізнаних ним пар мають подібність за іншими методами порівняння нижче рівня їх ефективної роботи, а 60-70% значень – нижче рівня їх застосовності. При чому відсоток помилково розпізнаних значень в таких межах не більше 30%, а вірно-позитивних близько 50%.

Реальний пошук пар буде обмежено блоками одного видавництва (отриманими з ISBN), а отже вплив хибно позитивних та невизначених варіантів буде незначним. Звісно, в значній частині блоків, які мали б належати одному видавництву, трапляються записи видавництв партнерів, але обмеження блоком все ж значно знижує відсоток хибних результатів.

Оскільки окрім скорочень існують і звичайні помилки введення, є наступні варіанти дій:

- об’єднати множини пар розпізнаних за допомогою:
- пропонованого методу, нечутливого до скорочень;
- одного з методів нечіткого порівняння в межах їх ефективності

пристосувати пропонований метод до невеликих помилок.

У першому варіанті необхідно поставити у відповідність парам, які знайдені пропонованим методом, «коефіцієнт подібності» в межах зони ефективного застосування методу, який застосовується на пару з ним.

Останній варіант краще з огляду на те, що методи нечіткого порівняння Яро-Вінклера та Левенштейна недостатньо інтелектуальні та можуть суттєво помилятися. Наступна пара має збіг на 96% за методом Яро-Вінклера, але є хибно-позитивним результатом:

- «УКРАІНСЬКА ВІЛЬНА АКАДЕМІЯ НАУК У США»;
- «УКР ВІЛЬНА АКАДЕМІЯ НАУК У КАНАДІ».

На такій великій кількості інформації пропонований метод не припускається помилок, а отже в разі пристосування його до роботи з даними, що мають одиночні помилки введення, він був би більш ефективним.

Висновки

Використання полів «Видавництво» та «ISBN» в задачі інтеграції гетерогенних бібліографічних даних є загально визнаним і продовжує досліджуватись. Методи, що застосовуються в ході такої інтеграції мають виявлену недосконаліть на назвах видавництв з скороченнями.

У роботі пропонується розроблений метод порівняння назв видавництв, нечутливий до скорочень. Використання такого методу в інтеграції даних дозволяє вирішити знайдену проблему. З огляду на специфіку, наявність 22% хибно-позитивних спрацювань методу не вплине на його застосовність до задач інтеграції бібліографічних даних.

60-85% знайдених за допомогою методу попарних збігів знаходяться за межами ефективного застосування методів порівняння, які використовуються в інтеграції даних загалом та бібліографічних даних зокрема. Отже цей метод доповнює існуючі методи та не може бути ними замінений. Описано сумісне застосування пропонованого та існуючих методів та можливі напрямки розвитку методу.

Отримані з використанням пропонованого методу пари варіантів назв одного і того ж видавництва можна використати для створення авторитетного файлу назв видавництв. Створений файл можна застосовувати для пошуку збігів на записах, які не мають ISBN а отже не можуть бути згрупованими за блоком видавництва та порівняні з достатньою точністю.

Також порівняння можна застосовувати в рамках пошуку дублікатів записів книг, враховуючи інші уточнюючі бібліографічні поля («Назва», «Автор», «Кількість сторінок» і т.д.), щоб мінімізувати вплив хибно-позитивних результатів порівняння.

Література

1. Online catalogs: what users and librarians want: an OCLC report. [Virtual Resource] – OCLC, 2009. – Access Mode: https://www.oclc.org/content/dam/oclc/reports/onlinecatalogs/fullreport.pdf.

2. Connaway, L. Publisher Names in Bibliographic Data: An Experimental Authority File and a Prototype Application, Library Resources and Technical Services [Text] / L. Connaway, T. Dickey. – OCLC, 2011. – №. 55.

3. Hahn, J. Instance mining: toward authoritative publisher entities using association rules [Virtual Resource] / J. Hahn // SWIB conference, – USA, 2020. – Access Mode: https://swib.org/swib20/slides/03-04-hahn.pdf.

4. Cohen, W. Integration of heterogeneous databases without common domains using queries based on textual similarity [Text] / W. Cohen // ACM SIGMOD-98. – 1998. –201-212. pp.

5. Bilenko, M. Adaptive Name Matching in Information Integration [Text] / M. Bilenko , R. Mooney, W. Cohen, P. Ravikumar, S. Fienberg // IEEE Intelligent Systems. – 2003. – Vol. 18. – Access Mode: https://www.cs.utexas.edu/~ml/papers/marlin-ieeeis-03.pdf.

6. Anish, R. Company name standardization using a fuzzy NLP approach [Virtual Resource] / R. Anish, G. Shashank, D. Paulami // Analytics Insight Magazine. – 2020. – Access Mode: https://www.analyticsinsight.net/company-names-standardization-using-a-fuzzy-nlp-approach/.

7. Paleo, B. A Modified Edit-Distance Algorithm for Record Linkage in a Database of Companies [Virtual Resource] / B. Paleo, C. Ghedini, J. Lima, C. Ribeiro, J. Filho // – 2006. – Access Mode: https://www.academia.edu/188363/A_Modified_Edit-Distance_Algorithm_for_Record_Linkage_in_a_Database_of_Companies.

8. Петренко, М. В. Проблема підвищення якості поля «Видавець» бібліотечних баз даних / М. В. Петренко // Modern directions of scientific research development : V international scientific and practical conference – Chicago : BoScience Publisher, 2021. –285–290 pp..

9. Taghva K. Recognizing Acronyms and their Definitions [Virtual Resource] / K. Taghva, J. Gilbreth// Technical Report 95-03, Information Science Research Institute. – UNLV, 1995. – Access Mode: http://www.isri.unlv.edu/ir/publications/Taghva95-03.ps.

10. Yeates, S. A. Automatic Extraction of Acronyms from Text / S. A. Yeates // New Zealand Computer Science Research Students Conference. – 1999. –117-124 pp..

11. Петренко, М. В. Дослідження візуально непомітних помилок введення та їхнього впливу на якість і пошукову доступність бібліографічних даних / М. В. Петренко // Реєстрація, зберігання і обробка даних, – Київ, 2021. – №23(2). – С. 81–90.

12. Петренко, М. В. Дослідження особливостей поля «ISBN» та їх впливу на процес обробки бібліографічних даних / М. В. Петренко // Вчені записки ТНУ імені В.І. Вернадського. Серія: Технічні науки, – Київ, 2021. – №32(4). – С.130–136.

13. Електронний каталог бібліотеки ім. Лесі Українки (публічні бібліотеки для дорослих м. Києва) [Електронний ресурс] – Режим доступу: http://ecatalog.kiev.ua.

14. Електронний каталог бібліотеки ім. Тараса Шевченка для дітей (публічні бібліотеки для дітей м. Києва) [Електронний ресурс] – Режим доступу: http://zra.kiev.ua:8081/MarcWeb.

References

1. Online catalogs: what users and librarians want: an OCLC report. [Virtual Resource]– OCLC, 2009. – Access Mode: https://www.oclc.org/content/dam/oclc/reports/onlinecatalogs/fullreport.pdf.

2. Connaway, L. Publisher Names in Bibliographic Data: An Experimental Authority File and a Prototype Application, Library Resources and Technical Services [/ L. Connaway, T. Dickey. – OCLC, 2011. – №. 55.

3. Hahn, J. (2020, November). Instance mining: toward authoritative publisher entities using association rules [Virtual Resource] / J. Hahn // SWIB conference, – USA, 2020. – Access Mode: https://swib.org/swib20/slides/03-04-hahn.pdf.

4. Cohen, W. Integration of heterogeneous databases without common domains using queries based on textual similarity / W. Cohen // ACM SIGMOD-98. – 1998. – pp. 201-212.

5. Bilenko, M. Adaptive Name Matching in Information Integration [Text] / M. Bilenko , R. Mooney, W. Cohen, P. Ravikumar, S. Fienberg // IEEE Intelligent Systems. – 2003. – Vol. 18. – Access Mode: https://www.cs.utexas.edu/~ml/papers/marlin-ieeeis-03.pdf.

6. Anish, R. Company name standardization using a fuzzy NLP approach [Virtual Resource] / R. Anish, G. Shashank, D. Paulami // Analytics Insight Magazine. – 2020. – Access Mode: https://www.analyticsinsight.net/company-names-standardization-using-a-fuzzy-nlp-approach/.

7. Paleo, B. A Modified Edit-Distance Algorithm for Record Linkage in a Database of Companies [Virtual Resource] / B. Paleo, C. Ghedini, J. Lima, C. Ribeiro, J. Filho // – 2006. – Access Mode: https://www.academia.edu/188363/A_Modified_Edit-Distance_Algorithm_for_Record_Linkage_in_a_Database_of_Companies.

8. Petrenko, M. V. Problema pidvischennya yakosti polya «Vidavec'» bibliotechnih baz danih / M. V. Petrenko // Modern directions of scientific research development : V international scientific and practical conference - Chicago : BoScience Publisher, 2021. - 285-290 pp..

9. Taghva K. Recognizing Acronyms and their Definitions [Virtual Resource] / K. Taghva, J. Gilbreth// Technical Report 95-03, Information Science Research Institute. – UNLV, 1995. – Access Mode: http://www.isri.unlv.edu/ir/publications/Taghva95-03.ps.

10.Yeates, S. A. Automatic Extraction of Acronyms from Text / S. A. Yeates // New Zealand Computer Science Research Students Conference. – 1999. – pp. 117-124.

11.Petrenko, M. V. Doslidzhennya vizual'no nepomitnih pomilok vvedennya ta ïhn'ogo vplivu na yakist' i poshukovu dostupnist' bibliografichnih danih / M. V. Petrenko // Reєstraciya, zberigannya i obrobka danih, - Kiïv, 2021. - №23(2). - S. 81-90.

12.Petrenko, M. V. Doslidzhennya osoblivostej polya «ISBN» ta ïh vplivu na proces obrobki bibliografichnih danih / M. V. Petrenko // Vcheni zapiski TNU imeni V.I. Vernads'kogo. Seriya: Tehnichni nauki, - Kiïv, 2021. - №32(4). - C.130-136.

13.Elektronnij katalog biblioteki im. Lesi Ukraïnki (publichni biblioteki dlya doroslih m. Kiєva) [Elektronnij resurs] - Rezhim dostupu: http://ecatalog.kiev.ua.

14.Elektronnij katalog biblioteki im. Tarasa Shevchenka dlya ditej (publichni biblioteki dlya ditej m. Kiєva) [Elektronnij resurs] - Rezhim dostupu: http://zra.kiev.ua:8081/MarcWeb.