АВТОМАТИЧНО ЗГЕНЕРОВАНІ КОДИ ЗА УНІВЕРСАЛЬНОЮ ДЕСЯТКОВОЮ КЛАСИФІКАЦЄЮ (УДК) ДЛЯ НЕКЛАСИФІКОВАНИХ ДОКУМЕНТІВ: РЕАЛЬНІСТЬ ЧИ ПЕРСПЕКТИВА

***Муравйова В.М.**, завідувач відділу класифікаційних систем,*
ДНУ «Книжкова палата України ім. Івана Федорова»

Універсальна десяткова класифікація є однією з найпоширеніших класифікаційних таблиць (особливо у Європі) для упорядкування всіх галузей знань. Вона використовується у бібліотеках, видавництвах, бібліографічних та інформаційних службах у понад 130 країнах світу та видається більш ніж 40 мовами, скорочені таблиці УДК доступні 57 мовами. Індекс УДК складається з арабських цифр та загальних розділових знаків. Саме тому він уможливлює ідентифікацію вітчизняної книжки або іншої публікації у базі даних будь-якої країни світу, що використовують УДК, оскільки не залежить від конкретної мови.

Міжнародним еталоном УДК є англомовна база даних Master Reference File (MRF), яка налічує близько 72 000 класифікаційних рубрик. Власником є Консорціум УДК (Нідерланди, м. Гаага).

У межах Ліцензійної угоди Книжкова палата України співпрацює із Консорціумом УДК, завдяки чому забезпечується безперервна актуалізація таблиць. В Україні з 2000 року використовують таблиці УДК українською мовою. З 2017 р. національний еталон УДК став єдиною класифікаційною системою.

З розвитком комп'ютерних технологій імплементувалася і робота бібліотек та видавничих організацій. Вони займаються такими питаннями, як організація, зберігання, обробка та класифікація документів. На їхніх сайтах стали доступними електронні документи та каталоги з урахуванням пошукових систем за різними критеріями.

Як зазначають автори статті [2] "Багато документів у бібліотеках по всьому світу все ще класифікуються вручну – або через брак довіри до автоматичної класифікації або через її недостатню ефективність.

Недовіра до автоматичної класифікації зрозуміла з точки зору бібліотекарів, оскільки неправильна класифікація в процесі каталогізації спричиняє багато додаткової роботи з редагуванням запису. Більше того, такий документ важко знайти".

З другого боку, систематизатори та редактори використовують автоматичні процеси в інших сферах життя та хочуть щоб і їхня робота була більш ефективною і полегшувалася за рахунок автоматичних підказок. Чи зможе штучний інтелект (ШІ) замінити людину чи допомогти при систематизації документів?

Книжка чи журнал починають своє життя з видавничої установи. Наскільки видавець відповідально відноситься до своєї роботи на стільки якісним продукт і вийде. Більшість видавців зневажають не тільки щодо проставляння правильного шифру видання чи індексу УДК у журналі, а і його оформлення. У результаті чого, цю книжку/статтю переіндексовують вручну в залежності від типу бібліотеки. Те ж саме стосується і журналів. Із-за таких помилок вже лунають звернення з закордону, що у нас "неправильне" УДК.

Можливо у подальшому автоматизований/напівавтоматизований процес і дозволить уникати грубих помилок при класифікації документів.

Наприклад, у статті Павловської Т. С. "Тестове оцінювання знань учнів із тематичного розділу шкільного курсу географії 8-го класу "Природні умови й ресурси України" // Природнича освіта та наука. – 2023. – №1. – (Видавничий дім "Гельветика") індекс УДК 502(477)(073) не відповідає пошуковому запиту користувача, який шукає точну інформацію по тематиці "Навколишнє середовище – Екологія – Україна – навчальний план", бо не відображає зміст статті. Тому такий індекс УДК є хибним. Натомість повинен бути такий індекс:

УДК 373.5**.091.279.7.*091.214***:913(477**)]**кл8***(045), де:
373.5 Базова та профільна середня освіта
37**.091.279.7** Виставлення оцінок. Оцінювання
37***.091.214*** Навчальні програми. Плани. Розклад занять. Плани лекцій,семінарів, консультацій,уроків, календарні плани
913 Регіональна географія
(477) Україна
*кл8 8 клас
(045) Статті в періодичних виданнях; …

Цей індекс складений із 7 частин, побудований за допомогою 2-х основних класів, 2-х спеціальних і 2-х загальних визначників та запозиченої нотації. І відповідає таким критеріям пошуку " Шкільний курс – географія – 8-й клас – оцінювання – навчальна програма, як предмет", що відповідає даній статті.

Чи зможе ШІ автоматично згенерувати даний код УДК чи лише дати складові індексів УДК, а систематизатор вже відкоригує індекс УДК. Стосовно автоматичного згенерованого індексу УДК маю сумніви, а от щодо напівавтоматичного вже є розроблені програми.

Автоматично допускаю, що можуть бути побудовані прості індекси, такі наприклад, як хімія – УДК 54, наука – УДК 001; можливий навіть варіант складних із 2-5 складників кодів УДК: географія **України** – УДК 913**(477)**, хімія: **підручник *для 10 кл*** – УДК 54***кл10*(075.3)**.

Автори статті [2] вивчивши структуру побудови та методики за класифікаційною системою УДК надають можливість гібридного підходу до побудованих індексів УДК під час каталогізації документів на основних рівнях без використання спеціальних визначників. І не виключають можливості в майбутньому "поекспериментувати з включенням методів колаборативної фільтрації та інших архітектур глибоких нейронних мереж, оскільки останні останнім часом досягли значних успіхів у сферах інтелектуального аналізу тексту та обробки природної мови".

Напівавтоматичний пошук та побудову індексів УДК вже можна отримати використовуючи УДК онлайн-хаб (https://udc-hub.com/). Він доступний лише для країн, які передплачують дану Ліцензію (плата за 1 рік становить €4500 якщо заплатити відразу за 4 роки, то – €16,200. Договір підписується на 12 років). На даний момент там відсутня наша національна версія УДК, але Книжкова палата України ще до повномасштабного вторгнення готова загрузити на сервер Консорціуму УДК українську версію таблиць.

У онлай-хабі доступний пошук за декількома критеріями:
– точної відповідності за номером УДК;
– точної відповідності за допомогою слів з можливістю його усічення;
– за фразою;
– використовуючи логічних операторів.
Із можливістю обмежити різними полями (доступно 16 полів).

Доступне і робоче місце систематизатора, де можна сконструювати індекс УДК. На цьому етапі розробки конструктор номерів підтримує дві функції:

а) запам'ятовувати числа, які потрібно зберегти, продовжуючи пошук/перегляд;

б) надати допомогу в об'єднанні чисел у складні рядки, готові для копіювання в локальну систему бази даних.

Інтерфейс тематично-алфавітного ланцюга індекса дає можливість влучно підібрати основний клас УДК.

Наприклад, слово "вода" може бути як і в хімії, так і водопостачанні або у водних ресурсах. Саме завдяки тематичному ланцюгу можна коректно підібрати індекс УДК, враховуючи всі складові статті.

Саме у складних випадках систематизатор краще зорієнтується чим ШІ. Інформації може бути забагато, але не основної. Тому не тільки ключові слова відіграють важливу роль при підборі індексу УДК. Важливий аналіз тематики, наприклад, у статті Є. Єльпітіфорова "Viscum album на рослинах, що можуть/не можуть бути господарями для напівпаразита – порівняльний мікроелементний склад" ключові слова: **омела (паразит-шкідник),** береза, вяз та тополя, мікроелементи мають такі складові індексу УДК

632.53 **шкідники** рослини для рослин
582.728.22 **омела**
582.62/.63 дерева
577.118 мікроелементи

Не дивлячись на те, що мова йде у першу чергу про омелу, як шкідника, перевагу надаємо **мікроелементам**:

577.118:[[632.53:582.728.22]:582.62/.63,

бо автор – біолог і пише саме із цієї точки зору, і потрапивши у сільське господарство (632.53) інформація для пошуку даної тематики була б незрозумілою.

Поки що Консорціумом УДК не активований інтерфейс "Перевірка/Розбір", який надасть змогу аналізувати індекси УДК.

У цілому ресурс корисний для напівавтоматичної побудови індексів УДК. Все залежить від обраної ліцензії.

Стосовно автоматичного генерування кодів УДК потрібно не забувати використання повністю ліцензованої версії класифікаційної системи УДК, оскільки це не порушує авторське право Консорціуму УДК, який є власником.

Список літератури

1. Книжкова палата України. http://www.ukrbook.net/UDC/UDC_1.html.
2. Універсальна десяткова класифікація. https://udcc.org/index.php.
3. Borovič, M., Ojsteršek, M., & Strnad, D. (2022). A Hybrid Approach to Recommending Universal Decimal Classification Codes for Cataloguing in Slovenian Digital Libraries. IEEE Access, (10), 85595-85605. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9856668.