Міністерство освіти і науки України
Харківський національний університет імені В. Н. Каразіна

Добровольський Геннадій Анатолійович

УДК 004.8:004.912:001.811

Модель, метод та інформаційна технологія відбору наукових публікацій у процесі підготовки бібліографічного покажчика

05.13.06 — інформаційні технології

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків — 2021

Дисертацією є рукопис.

Робота виконана в Запорізькому національному університеті Міністерства освіти і науки України.

Науковий керівник: кандидат фізико-математичних наук, доцент
Єрмолаєв Вадим Анатолійович,
Запорізький національний університет, доцент кафедри комп'ютерних наук.

Офіційні опоненти: Доктор технічних наук, доцент
Глибовець Андрій Миколайович,
Національний університет «Києво-Могилянська академія», декан факультету інформатики;

Доктор технічних наук, професор
Машталір Сергій Володимирович,
Харківський національний університет радіоелектроніки, професор кафедри інформатики.

Захист відбудеться «12» травня 2021 р. о 15^{15} на засіданні спеціалізованої вченої ради Д 64.051.09 Харківського національного університету імені В. Н. Каразіна за адресою: 61022 м. Харків, майдан Свободи, 6, ауд. 3-18.

З дисертацією можна ознайомитись у Центральній науковій бібліотеці Харківського національного університету імені В.Н.Каразіна за адресою: 61022, м.Харків, майдан Свободи,4.

Автореферат розісланий «08» квітня 2021 р.

Учений секретар
спеціалізованої вченої ради Олена ТОЛСТОЛУЗЬКА

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. У процесі виконання наукових досліджень, що включає аналіз останніх досягнень, вивчення нових тенденцій, підтримку обізнаності, пошук рецензентів та колег для спільних проектів, потрібно знайти та проаналізувати публікації, які містять всі важливі наукові результати – створити науково-допоміжний або рекомендаційний бібліографічний покажчик на задану тему. Етапами створення бібліографічного покажчика є бібліографічне виявлення, та відбір. Бібліографічне виявлення – це складання переліку публікацій, які можуть після відбору увійти до бібліографічного покажчика.

Попередній відбір відбувається одночасно з виявленням публікацій під час їх пошуку в архівах журналів, конференцій, використання наукових пошукових систем таких як Google Scholar, Scopus, Microsoft Academic. У межах існуючих практик попередній відбір здійснює експерт шляхом перегляду публікацій з максимальною повнотою, щоб вирішити, чи потрібно їх включити до покажчика. Остаточний відбір відбувається після систематизації всього зібраного матеріалу та виявлення прогалин або нерелевантних публікацій. Складання бібліографічного покажчика – процес трудомісткий. Тому на практиці бібліотеки на запити окремих користувачів укладають бібліографічні довідки (списки літератури), але не складають бібліографічні покажчики. Крім того, це, зазвичай, робиться бібліотечними робітниками, які не мають необхідного експертного досвіду у заданій галузі досліджень.

Відомі методи аналізу наукових публікацій включають інтелектуальний аналіз текстів (S. Deerwester, S. Dumais, G. Furnas, T. Landauer, D. M. Blei, X. Yan, Y. Zuo, Д. В. Ланде, Н. В. Шаронова, Н. Ф. Хайрова, А. М. Глибовець, К. В. Воронцов та ін.), дослідження цитувань та методи обчислення наукової цінності публікацій (D. J. de Solla Price, E. Garfield, L. Leydesdorff, P. Wouters, M. Callon, J. Law, A. Rip, R. Merton, J. R. Cole, S. Cole, J. Hirsch, M. Newman, D. Edge, J. C. Shin, L. Egghe, N. Hummon, P. Dereian, B. Батагель, J. S. Liu, L. YY. Lu, C. Calero-Medina, E. Noyons, J. Lecy. Д. В. Ланде, А. М. Глибовець, Н. Ф. Хайрова, В. Д. Гогунський та ін.), методи створення систематичних оглядів стану досліджень у вибраній науковій області (M. Petticrew, S. Gilbody та ін.). Аналіз існуючих досліджень показав, що вони або зосереджуються на вивченні вже готових корпусів публікацій, або виконують пошук за ключовими словами для збору таких корпусів. Суттєвими недоліками існуючого способу пошуку є його

суб'єктивність, нестача систематичності, брак об'єктивних критеріїв повноти та зупинки пошуку. Відсутність досвіду у заданій області, індивідуальні особливості термінології, наукові інтереси та досвід конкретної особи спричиняють появу нерелевантних публікацій, прогалин у результуючому корпусі. В існуючих методах та практиках покращення якості пошуку виконується за допомогою корекції набору ключових слів шляхом вивчення всіх виявлених публікацій – як релевантних, так і нерелевантних, – що спричиняє зайві витрати часу і не гарантує повноти корпусу. На підставі аналізу зроблено висновок про те, що актуальною є науково-практична задача підвищення ефективності процесу бібліографічного виявлення й відбору шляхом розробки удосконалених моделі, методу та інформаційної технології.

Зв'язок роботи з науковими програмами, планами, темами Дисертаційне дослідження виконувалось в рамках міжнародного проекту SemData (грант PIRSES-GA-2013-612551 програми стипендій для дослідників Marie Sklodowska-Curie Actions) та НДР МОН України «Методи та технології інформаційного пошуку за схожістю та зіставлення записів» (№ДР 0113U001936) кафедри інформаційних технологій Запорізького національного університету.

Мета і завдання дослідження. Метою дисертаційної роботи є підвищення ефективності виявлення та відбору наукових публікацій у процесі підготовки науково-допоміжного або рекомендаційного бібліографічного покажчика шляхом отримання мінімальної термінологічно насиченої впорядкованої множини публікацій за допомогою розробки та застосування інформаційної технології, яка базується на розроблених математичній моделі та методі бібліографічного виявлення й відбору, та відповідні програмні рішення.

Для досягнення мети в роботі необхідно вирішити такі завдання:

– проаналізувати наявний процес бібліографічного виявлення та відбору з метою визначення його недоліків;

– виявити моделі, методи та інформаційні технології, які можуть бути вдосконалені, об'єднані та застосовані для підвищення ефективності процесу бібліографічного виявлення та відбору;

– розробити гібридну математичну модель процесу бібліографічного виявлення та відбору, призначену для представлення, ітеративного виявлення та відбору мінімальної термінологічно насиченої впорядкованої множини публікацій;

– розробити метод бібліографічного виявлення та відбору мінімальної термінологічно насиченої впорядкованої множини публікацій;

– розробити інформаційну технологію бібліографічного виявлення та

відбору, що базується на розроблених моделі та методі;
– програмно реалізувати розроблену інформаційну технологію;
– провести експериментальне дослідження ефективності розробленої інформаційної технології;
– провести практичну апробацію розробленої інформаційної технології.

Об'єктом дослідження є процес підготовки науково-допоміжних та рекомендаційних бібліографічних покажчиків у наукометричних пошукових системах.

Предметом дослідження є модель, метод та інформаційна технологія автоматизації бібліографічного виявлення та відбору з використанням ймовірнісного тематичного моделювання текстових документів, аналізу цитувань в наукометричних пошукових системах та автоматичного виявлення термінів.

Методи дослідження базуються на застосуванні: апарату теорії множин і методів аналізу графів для опису розроблених моделі та методу; алгоритмів машинного навчання і обробки природних мов для тематичного моделювання та автоматичного виявлення термінів; методів аналізу графів для виявлення й відбору найвпливовіших публікацій, математичної статистики для обчислення критеріїв зупинки бібліографічного виявлення та відбору.

Наукова новизна одержаних результатів:

– вперше розроблено гібридну математичну модель процесу бібліографічного виявлення та відбору, яка відрізняється від наявних одночасним використанням ітерацій контрольованої «снігової кулі», ймовірнісного тематичного моделювання текстових документів, аналізу мережі цитування, автоматичного виявлення термінів, наявністю ознак нерухомої точки ітерацій та показників насиченості набору термінів обраної предметної області, що створило підґрунтя для розробки методу виявлення та відбору мінімальної термінологічно насиченої впорядкованої множини публікацій;

– вперше розроблено метод бібліографічного виявлення та відбору, який відрізняється від наявних аналогів спільним узгодженим застосуванням вдосконалених методів відбору початкової множини публікацій, побудови ймовірнісної тематичної моделі текстових документів, контрольованої «снігової кулі», аналізу мережі цитування, автоматичного виявлення термінів та виявлення термінологічного насичення. На відміну від відомих аналогів розроблений метод дозволяє у процесі бібліографічного виявлення та відбору швидше отримати мінімальну термінологічно насичену впорядковану множину публікацій;

– вперше розроблено інформаційну технологію, що, на відміну від

відомих аналогів, дозволяє створювати мінімальні термінологічно насичені впорядковані множини публікацій, формуючи короткі бібліографічні покажчики, які містять найважливіші терміни предметної області;

– удосконалено метод імовірнісного тематичного моделювання текстових документів шляхом застосування методу головних компонент, що, на відміну від відомих аналогів, дозволяє автоматично визначити кількість тем;

– отримали подальший розвиток методи інформаційного пошуку, призначені для задоволення інформаційної потреби з низькою специфічністю, шляхом розробки критерію зупинки пошуку, що дозволяє сформувати мінімальну термінологічно насичену впорядковану множину публікацій і тим самим скоротити час пошуку.

Практичне значення одержаних результатів визначається тим, що розроблені модель, метод та інформаційна технологія можуть бути ефективно використані для автоматичного збору та підтримання в актуальному стані персоналізованих бібліографічних покажчиків – колекцій публікацій, що містять всі важливі результати заданої області наукового знання і, одночасно, є достатньо короткими для детального вивчення науковцем. Створені за допомогою розробленої технології бібліографічні покажчики можуть бути основою для аналізу сучасного стану досліджень – обов'язкової частини наукової роботи. У навчальному процесі розроблена технологія може служити засобом складання робочих програм дисциплін спеціалізації, переліку рекомендованих для вивчення публікацій. Розроблена інформаційна технологія і розроблені для її підтримки інструментальні програмні засоби були успішно впроваджені в навчальному процесі Українського католицького університету в дисципліні «Automated Term Extraction and Ontology Learning from Texts», яка викладається магістрантам спеціальності «Комп'ютерні науки» (акт про впровадження від 20.06.2018), та Запорізького національного університету при написанні кваліфікаційних робіт бакалаврів та магістрів(акт про впровадження від 7.10.2019).

Публікації. Основні результати дисертації опубліковані у 21 науковій праці, з них: 3 статті у наукових фахових виданнях України [1, 2, 3]; 5 наукових праць у зарубіжних спеціалізованих виданнях, проіндексованих в міжнародній науко-метричній базі Scopus [4, 5, 6, 7, 8]; 11 наукових праць, які засвідчують апробацію матеріалів дисертації [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19], 2 публікації, що додатково відображають наукові результати дисертації [20, 21], у тому числі 1 патент [21].

Особистий внесок здобувача. Усі наукові результати, викладені в дисертації, отримані здобувачем особисто. У роботах, які написано у співавторстві, особистий внесок здобувача полягає у наступному: сформульовано та реалізовано [9] метод визначення схожості коротких текстів за допомогою нейронної мережі; виконано обчислювальні експерименти з порівняння методів визначення схожості коротких текстів [20]; запропоновано один із варіантів визначення схожості коротких текстів шляхом порівняння їх хешів на основі сигнатур та їх пошуку [10]; метод порівняння коротких текстів за допомогою їх хешів на основі сигнатур застосовано до практичної задачі очистки даних [11]; виконано експерименти з метою визначення кількісних показників якості пошуку за допомогою хешів на основі сигнатур [12]; сформульовано математичну модель експерименту із визначення ймовірності колізій запропонованої хеш-функції [13]; отримано патент на корисну модель [21]; запропонований метод порівняння коротких текстів та пошуку реалізовано у вигляді програмної бібліотеки [14]; реалізовано покращений варіант методу порівняння та пошуку коротких текстів [1]; запропоновано архітектуру інформаційної системи на основі потоків даних [15, 18]; запропоновано удосконалення тематичної моделі текстових документів за допомогою методу головних компонент та її застосування до пошуку наукових публікацій [4]; розроблено вдосконалений за допомогою додаткової вимоги розрідженості метод невід'ємної матричної факторизації, який використовується для реалізації методу головних компонент та тематичного моделювання текстових документів [16]; розроблено ітеративний метод контрольованої «снігової кулі» для пошуку наукових публікацій реалізовано програмну бібліотеку для аналізу головних шляхів [19]; на основі розробленого методу бібліографічного виявлення та відбору розроблено інформаційну технологію [5]; розроблену інформаційну технологію застосовано до підготовки бібліографічного покажчика [2]; показано збіжність ітерацій розробленого методу контрольованої «снігової кулі» [6]; розроблено модель виявлення та відбору наукових публікацій [3]; вдосконалено метод автоматичного виявлення термінів [7, 8].

Апробація результатів дисертації. Основні положення та результати досліджень пройшли апробацію та отримали позитивну оцінку під час очних доповідей на наступних конференціях:

– X міжнародна конференція з математичного моделювання, МКММ 2009, смт. Залізний Порт (Херсонська обл.), 14-19 вересня 2009 р.;
– XI міжнародна конференція з математичного моделювання, МКММ 2010, смт. Залізний Порт (Херсонська обл.), 13-18 вересня 2010 р.;
– XII міжнародна конференція з математичного моделювання, МКММ 2011, смт. Залізний Порт (Херсонська обл.), 12-17 вересня 2011 р.;
– XIII міжнародна конференція з математичного моделювання, МКММ 2012, смт. Залізний Порт (Херсонська обл.), 17-22 вересня 2012 р.;
– International Conference on ICT in Education, Research, and Industrial Applications: Integration, Harmonization, and Knowledge Transfer, ICTERI 2012, м. Херсон, Херсонський державний університет, 6-10 червня 2012 р.;
– VIII всеукраїнська, XV регіональна наукова конференція молодих дослідників «Актуальні проблеми математики та інформатики», м. Запоріжжя, ЗНУ, 27-28 квітня 2017 р.;
– International Conference on ICT in Education, Research, and Industrial Applications: Integration, Harmonization, and Knowledge Transfer, ICTERI 2017, м. Київ, Київський національний університет ім. Тараса Шевченка, 15-18 травня 2017 р.;
– The 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2017, м. Бухарест, Румунія, 21-23 вересня 2017 р.;
– International Conference on Knowledge Engineering and Semantic Web, KESW 2017, м. Щецин, Польща, 08-10 листопада 2017 р.;
– IX всеукраїнська, XVI регіональна наукова конференція молодих дослідників «Актуальні проблеми математики та інформатики», м. Запоріжжя, ЗНУ, 26-27 квітня 2018 р.;
– International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer, ICTERI 2018, м. Київ, Київський національний університет ім. Тараса Шевченка, 14-17 травня 2018 р.;
– XIX міжнародна конференція з математичного моделювання, МКММ 2018, смт. Лазурне, Херсонська область, 17-21 вересня 2018 р.;
– International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. ICTERI 2019, м. Херсон, Херсонський державний університет, 12-15 травня 2019 р.

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, 4 розділів, загальних висновків, списку використаних джерел та 4 додатків. Обсяг загального тексту дисертації складає 216 сторінок, з

них основного тексту 156 сторінок, Робота ілюстрована 13 таблицями та 25 рисунками. Список використаних джерел містить 204 найменування.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** наведена загальна характеристика роботи: обґрунтована актуальність теми дисертаційної роботи, сформульовані мета і завдання дослідження, визначені об'єкт, предмет і методи дослідження, наукова новизна та практичне значення отриманих результатів, приведені дані про публікації, особистий внесок здобувача в роботи, виконані в співавторстві, та дані про апробацію результатів дисертації.

У **першому розділі** проаналізовано наявний технологічний процес підготовки бібліографічного покажчика з метою окреслення підпроцесів, які можуть бути автоматизовані, та методів, якими автоматизація може бути виконана. Визначено, що автоматизації засобами інформаційних технологій піддаються бібліографічне виявлення, відбір та групування, але для збору публікацій, що містять всі важливі наукові результати деякої області досліджень потрібно виконувати найскладніший вид виявлення, відбору та групування – рекомендаційне. Водночас кількість читачів бібліографічного покажчика, складеного для однієї або кількох публікацій, є занадто малою, тому бібліотеки зазвичай не виконують таку роботу.

Розглянуто особливості наукових публікацій та сучасний стан методів, які після вдосконалення можуть стати частиною інформаційної технології, призначеної для рекомендаційного бібліографічного виявлення та відбору: пошук за ключовими словами, рекомендаційні системи та наукометричні методи обчислення наукової цінності публікацій, особлива увага приділена методам створення систематичних оглядів літератури, аналізу мереж цитування та способам їх побудови методом «снігової кулі» та мірам схожості коротких текстів.

Аналіз показав, що найперспективнішою комбінацією є створення мережі цитування методом «снігової кулі», поєднане з тематичним моделюванням, аналізом цитування та автоматичним виявленням термінів. Фільтрація коротких текстів за допомогою ймовірнісного тематичного моделювання допоможе порівняти їх зміст, зменшивши тим самим розмір «снігової кулі», аналіз цитування дасть можливість відібрати визнані науковою спільнотою публікації, методи автоматичного виявлення термінів дозволять оцінити термінологічну насиченість впорядкованої множини публікацій та сформувати бібліографічний покажчик, який може бути вивчений експертом за

мінімальний час. Відповідним критерієм ефективності є кількість елементів у мінімальній термінологічно насиченій впорядкованій множині публікацій.

У другому розділі розроблено гібридну математичну модель процесу бібліографічного виявлення та відбору. Процес бібліографічного виявлення та відбору – це процес вибору із множини всіх доступних для аналізу публікацій $\mathbb{D}$ деякої підмножини $\mathbb{B} \subseteq \mathbb{D}$, елементи якої задовольняють інформаційну потребу користувача бібліографічного покажчика.

Враховуючи багатоваріантність можливого формулювання інформаційної потреби та реалізації процесу виявлення та відбору, для досягнення мети дослідження, у роботі запропоновано підхід до формального опису процесу бібліографічного виявлення та відбору і сформульовані такі припущення:

Припущення 1. Інформаційна потреба складається із кількох неформальних вимог: належності всіх публікацій із $\mathbb{B}$ до заданої предметної області, їх важливості у заданій предметній області, таку їх невелику кількість, що дозволяє детальне вивчення користувачем, і наявність у них усіх основних термінів предметної області.

Припущення 2. Кожну публікацію $d \in \mathbb{D}$ можна відобразити у множину речень $\mathbb{S}(d)$, і кожне речення $s \in \mathbb{S}(d)$ – у множину словосполучень $\mathbb{C}(s)$, яка є підмножиною множини $\mathbb{C}$ всіх словосполучень, які зустрічаються у $\mathbb{D}$, де словосполученням $c \in \mathbb{C}$ називається слово або стійкий кортеж слів. Терміни $\mathbb{T}$ позначають поняття заданої предметної області і є підмножиною словосполучень $\mathbb{T} \subseteq \mathbb{C}$.

Припущення 3. Існує відображення цитування, визначене на множині $\mathbb{D}$ доступних для аналізу публікацій:

$$REF : \{v\} \to \{u \in \mathbb{D} \,|\, v \text{ цитує } u\},\ v \in \mathbb{D}. \tag{1}$$

Обернене відображення цитування визначається як

$$REF^{-1} : \{u\} \to \{v \in \mathbb{D} \,|\, v \text{ цитує } u\},\ u \in \mathbb{D} \tag{2}$$

і повторне відображення цитування:

$$REF^k = \begin{cases} REF & , k = 1, \\ REF \circ REF^{k-1} & , k > 1 \wedge k \in \mathbb{N} \end{cases} \tag{3}$$

Відображення (1) визначає орієнтований граф - мережу цитування

$$N = (\mathbb{D}, \mathbb{E}) \tag{4}$$

з дугами $\mathbb{E} = \{vu, \forall v \in \mathbb{D}, u \in \mathit{REF}(\{v\})\}$ та вершинами $d \in \mathbb{D}$.

Припущення 4. Мережа цитування (4) є майже ациклічною (В. Батагель):

$$\left|\left\{d \in \mathbb{D} \,|\, \exists k \in \mathbb{N}, d \in \mathit{REF}^k(\{d\})\right\}\right| \ll \left|\left\{d \in \mathbb{D} \,|\, \forall k \in \mathbb{N}, d \notin \mathit{REF}^k(\{d\})\right\}\right|, \tag{5}$$

де $k \in \mathbb{N}$ – довжина шляху в мережі цитування.

Припущення 5. Необхідною ознакою наявності у $\mathbb{B}$ всіх основних термінів предметної області є термінологічне насичення впорядкованої множини публікацій (В. А. Єрмолаєв, О. Татарінцева).

Припущення 6. Повний текст публікації недоступний. Часто обмеження, встановлені власниками авторських прав, ускладнюють доступ до повного тексту публікації, тому в розробленій інформаційній технології її повний текст застосовується лише на останніх етапах.

Розроблена формальна гібридна математична модель процесу бібліографічного виявлення та відбору визначається кортежем

$$\mathbb{M} = \langle \mathbb{D}, \mathit{REF}, \mathit{PTM}, \mathit{DocDiff}, \delta, \mathit{Snowball}, \mathbb{B}_0, \mathit{DocListDiff}, \omega, \mathit{SPC}, \mathit{MaxRank}, \mathit{Terms}, \mathit{Cvalue}, \mathit{thd} \rangle, \tag{6}$$

де $\mathbb{D}$ - доступні для аналізу публікації; *REF* - відображення цитування; *PTM* - представлення змісту публікації; $\mathit{DocDiff}$ - міра відмінності публікацій; δ - гранична міра відмінності публікацій; *Snowball* - відображення ітерацій «снігової кулі»; $\mathbb{B}_0$ - початкова точка ітерацій «снігової кулі»; *SPC* - вага публікації у предметній області; *DocListDiff* - міра близькості впорядкованих множин публікацій; ω - межа міри близькості впорядкованих множин публікацій; *MaxRank* - максимальний ранг публікацій; Terms - відображення публікацій $\mathbb{D}$ у множину термінів $\mathbb{T}$; $\mathit{Cvalue}(\tau)$ - вага терміна τ; thd - міра відмінності множин термінів.

Опис предметної області у моделі задається множиною публікацій $\mathbb{B}_0$ ($\mathbb{B}_0 \subseteq \mathbb{D}$, $|\mathbb{B}_0| \sim O(10)$), яка одночасно є початковою точкою ітерацій «снігової кулі».

Міра належності публікації до предметної області обчислюється за допомогою імовірнісної тематичної моделі текстових документів (ІТМ).

ІТМ представляє зміст кожної публікації $d \in \mathbb{D}$ у вигляді умовних імовірностей

$$p(t|d) = \mathit{PTM}\,(d)\,, \tag{7}$$

що показують імовірність належності публікації d до теми t. Кожна тема t визначається ймовірностями $p(\tau_i|t)$ належності словосполучення τ_i до теми t і апріорною ймовірністю $p(t)$.

У моделі використовується вдосконалена ІТМ, яка будується шляхом визначення розподілів $p(\tau_i|t)$ та $p(t)$, виходячи з частоти сусідства словосполучень

$$p(\tau_i, \tau_k) = \sum_t p(\tau_i|t)p(t)p(\tau_k|t), \tag{8}$$

яка підраховується як кількість речень s, у яких одночасно зустрілися τ_i та τ_k.

Запропонованим вдосконаленням ІТМ є застосування розрідженої невід'ємної симетричної матричної факторизації (РСНМФ) для тематичного моделювання. За допомогою визначення матриць A та H як
$A_{ij} = A_{ji} = p(\tau_i, \tau_j)$ та $H_{it} = p(\tau_i|t)\sqrt{p(t)}$, вираз для $p(\tau_i, \tau_k)$ перетворюється на $A = HH^T$, і задача тематичного моделювання зводиться до мінімізації модифікованої цільової функції

$$\left\|A - HH^T\right\|_F^2 + \lambda \sum_{it} |H_{it}| \to min, H_{ij} \geq 0. \tag{9}$$

Параметр λ впливає на ступінь розрідженості та похибку факторизації. Показано, що запроваджена регуляризація за допомогою доданка
$\lambda \sum_{it} |H_{it}|$ у поєднанні із симетричністю матриці A дозволяє визначити кількість тем аналогічно до методу головних компонент.

Представлення публікацій у вигляді набору ймовірностей дозволяє використати як міру відмінності публікацій $DocDiff$ один із найпоширеніших способів порівняння тематичних моделей текстових документів – розходження Кульбака-Лейблера, симетричне розходження Кульбака-Лейблера, відстань Гелінгера та розходження Дженсена-Шеннона.

Граничне значення міри відмінності δ експериментально вибрано таким, щоб на початку ітерацій обмеженої «снігової кулі» (10) включати до множини $\mathbb{B}_{i+1}$ близько 30% виявлених публікацій.

Відображення ітерацій «снігової кулі», призначене для бібліографічного виявлення, визначається як:

$$\begin{aligned}\mathbb{B}_{i+1} &= \mathit{Snowball}\left(\mathbb{B}_i\right) \\ &= \bigcup_{v \in \mathbb{B}_i} \{v\} \cup \mathit{REF}(\{v\}) \cup \mathit{REF}^{-1}(\{v\})\big|_{DocDiff(v, \mathbb{B}_0) < \delta},\end{aligned} \tag{10}$$

де $\mathbb{B}_i \subseteq \mathbb{D}$. Рівняння (10) відрізняється від аналогів (A. Ahad, M.Fayaz, A.S.Shah; J.Lecy, K.Beatty) використанням тематичної моделі текстових

документів для обчислення відмінності публікацій та переходами як за цитуваннями, так і в протилежному напрямку.

Вага публікації у предметній області

$$SPC_i : v \ \rightarrow \ \mathbb{N}, v \in \ \mathbb{B}_i, \tag{11}$$

визначається шляхом аналізу головних шляхів пошуку в підграфі $N_i \in N$ мережі цитування (4), побудованому із вершин $d \in \mathbb{B}_i$ та дуг $\mathbb{E} = \{vu, \forall v \in \mathbb{B}_i, u \in \mathbb{B}_i \cap REF(\{v\})\}$ після трансформації циклів в ациклічні фрагменти за допомогою препринтного перетворення (В. Батагель). SPC_i дозволяє знайти ранг $Rank_i(v)$ кожної публікації та визначити шукану впорядковану множину публікацій:

$$\begin{aligned} \mathbb{L}_i\left(MaxRank\right) = (v_k)_{k=1}^{|\mathbb{B}_i|}, & Rank_i(v_k) < MaxRank, \\ & Rank_i(v_k) \leq Rank_i(v_{k+1}), \end{aligned} \tag{12}$$

де максимальний ранг публікацій $MaxRank$ обмежує кількість публікацій у бібліографічному покажчику і визначається вимогою досягнення нерухомої точки ітерацій (10) та наявністю термінологічного насичення.

В рамках розробленої моделі мірою близькості впорядкованих множин публікацій $DocListDiff$ вибрана рангова кореляція Спірмена $\rho(\mathbb{L}_i, \mathbb{L}_{i+1})$, і умовою нерухомої точки ітерацій (10) є нерівність

$$|\rho(\mathbb{L}_i, \mathbb{L}_{i+1}) - 1| < \omega, i > i_0, \tag{13}$$

де ω – межа міри близькості впорядкованих множин публікацій (12), параметр, який задає рівень варіативності впорядкованої множини публікацій.

Відображення публікацій $\mathbb{L}_i$ у множину термінів $\mathbb{T}_i$

$$\mathbb{T}_i = Terms(\mathbb{L}_i) \tag{14}$$

відбувається шляхом застосування до об'єднаного тексту публікацій процедури автоматичного визначення термінів, запропонованої K. Frantzi, S. Ananiadou та H. Mima та вдосконаленої в роботах В. А. Єрмолаєва та ін., яка визначає вагу терміна $Cvalue_i(\tau)$ у множині публікацій $\mathbb{L}_i(MaxRank)$, граничне значення ϵ_i ваги терміна, та міру відмінності множин термінів $thd(\mathbb{T}_i, \mathbb{T}_j)$.

Термінологічне насичення впорядкованої множини публікацій, визначається вимогою, що додавання Δ публікацій у кінець переліку

(12) майже не змінює перелік термінів

$$\frac{thd(\mathbb{T}_i(MaxRank), \mathbb{T}_i(MaxRank + \Delta))}{\epsilon_i} < 1, \Delta > 0. \tag{15}$$

Мінімальна термінологічно насичена впорядкована множина публікацій визначається рівнянням (12), у якому

$$MaxRank = min\left\{ M \,\middle|\, \frac{thd(\mathbb{T}_i(M), \mathbb{T}_i(M + \Delta))}{\epsilon_i} < 1 \right\}. \tag{16}$$

Мірою ефективності розробленої моделі (6) є кількість публікацій $|\mathbb{L}_i|$ у бібліографічному покажчику, визначеному умовами (12), (13), (15) та (16).

У **третьому розділі** на основі розробленої моделі розроблено метод бібліографічного виявлення та відбору та інформаційну технологію, що надають можливість формування рекомендаційних та науково-виробничих бібліографічних покажчиків, а також оцінки їх необхідного розміру. Інформаційну технологію, що спирається на розроблений метод, представлено за допомогою моделі «4 + 1 точка зору». Опис технології включає її призначення, сценарій роботи, варіанти використання, структури даних та їх призначення, процес виявлення та відбору, компоненти. Діаграма діяльності UML для технології наведена на рисунку 1.

Описано розроблений метод бібліографічного виявлення та відбору, який містить наступні пов'язані між собою частини:

1. Вдосконалений метод відбору початкової множини публікацій, що відрізняється від аналогів зміненими критеріями ручного відбору.

2. Вдосконалений метод побудови ймовірнісної тематичної моделі текстових документів, що відрізняється від найближчих аналогів застосуванням методу головних компонент для визначення кількості тем.

3. Вдосконалений метод ітерацій контрольованої «снігової кулі», що відрізняється від аналогів застосуванням ймовірнісної тематичної моделі текстових документів та наявністю критерію нерухомої точки.

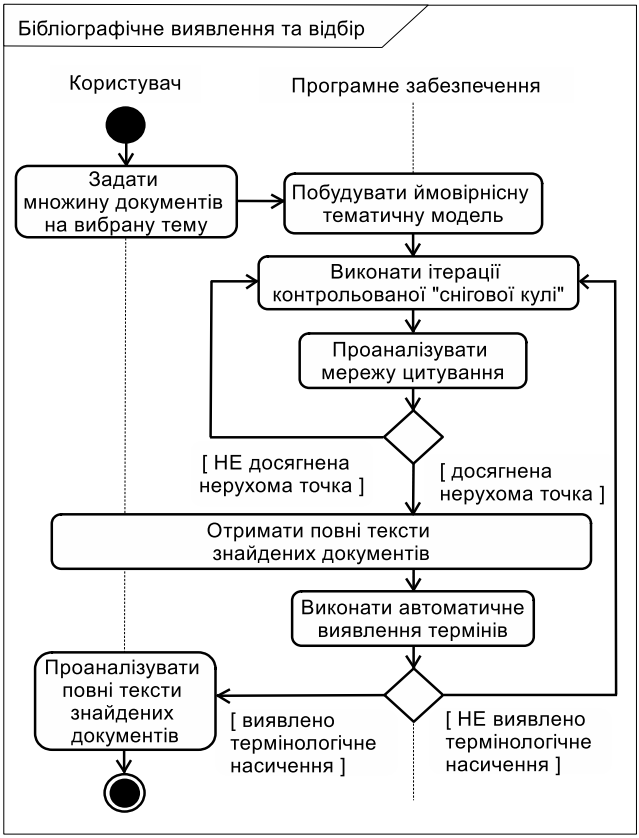

Рис. 1 Діаграма діяльності інформаційної технології бібліографічного виявлення та відбору

4. Вдосконалений метод аналізу мережі цитування, що відрізняється покращеним способом препринтного перетворення, який враховує перетин циклів у мережі цитування.

5. Найпростіший метод отримання повних текстів документів – пошук серед вільно доступних публікацій.

6. Вдосконалений метод автоматичного виявлення термінів, який відрізняється від попередника застосуванням скінченого автомата Ахо-Корасік для виявлення входжень коротких термінів у довгі.

7. Метод виявлення термінологічного насичення.

Кожен із названих пунктів є реалізацією відповідної частини розробленої інформаційної технології та спирається на розроблену раніше модель, описуючи деталі, необхідні для практичного застосування.

Міра ефективності інформаційної технології визначається, як розмір мінімальної термінологічно насиченої множини публікацій.

Четвертий розділ дисертації представляє мету, завдання, план проведення, методи виконання, методи вимірювання та аналіз результатів обчислювальних експериментів, виконаних із метою дослідження розробленого методу бібліографічного виявлення та відбору. Мережі цитування, які використовувалися в дисертації для аналізу, методи їх збору та коротку характеристику кожної колекції, яка показана в таблиці 1. Крім названих, для перевірки якості

Таблиця 1

Перелік використаних в експериментах мереж цитування

Мережа	**Тема**	**Міра відмінності**	**Кількість публікацій**
CAPT-KL	Методи комп'ютерного навчання вимові	Кульбака-Лейблера	6361
CAPT-SKL	Методи комп'ютерного навчання вимові	Симетрична Кульбака-Лейблера	4684
CAPT-JS	Методи комп'ютерного навчання вимові	Дженсена-Шеннона	6369
CAPT-HL	Методи комп'ютерного навчання вимові	Гелінгера	6862
ONTO-KL	Онтології	Кульбака-Лейблера	1068

тематичної моделі текстових документів використовувалась відома колекція публікацій HEP (High Energy Physics). Досліджено властивості початкової колекції публікацій: аналіз залежності індексу цитування публікації від її віку та аналіз відібраних публікацій, які є ідеальною початковою множиною. Показано, що критеріями відбору початкової колекції є кількість цитувань у релевантних публікаціях та рік публікації. Виконано оцінки якості вдосконаленої тематичної моделі текстових документів. Показано зв'язок розходження

Кульбака-Лейблера $D_{KL}(u,v)$ із відомою косинусною мірою схожості $cos\,(\theta(u,v))$

$$cos\,(\theta(u,v)) \leq 1 - D_{KL}(u,v)/3, \tag{17}$$

який перевірявся на колекції публікацій HEP. Також демонструється можливість автоматичного визначення кількості тем методом головних компонент. Досліджено мережу цитування (4), отриману методом контрольованої «снігової кулі». Показано, що мережа цитування (4) є «малим світом» – розрідженим зв'язним графом із однією гігантською компонентою та коротким середнім шляхом $\overline{Path(u,v)} \ll |\mathbb{B}_i|$, між вершинами графа (4) який прямо пропорційний середній кількості ітерацій методу контрольованої «снігової кулі». Досліджується збіжність ітерацій методу контрольованої «снігової кулі». Рисунок 2 демонструє наближення ітерацій до нерухомої точки для послідовності впорядкованих множин публікацій, отриманих для різних мереж цитування. Показано, що мережа на основі міри Кульбака-Лейблера забезпечує швидше наближення до нерухомої точки $\rho = 1$. Вивчається термінологічне насичення для переліку публікацій із теми «ontologies» (комп'ютерні науки), виявлених та відібраних розробленим у дисертаційній роботі методом (мережа цитування ONTO-KL).

Для порівняння отримуються та досліджуються переліки термінів із публікацій, які містяться в різних пошукових системах. Перший перелік термінів одержано з публікацій, які проіндексовані пошуковою системою «Microsoft Academic Search», мають у ній автоматично призначену категорію «ontologies» та впорядковані за індексом цитування. Другий перелік отримано з публікацій, які зберігаються в електронній бібліотеці «ACM digital library», мають присвоєну авторами мітку «ontologies» та впорядковані за індексом цитування. Третій перелік термінів отримано з публікацій, знайдених у пошуковій системі «Google Академія» за ключовим словом «ontologies». Термінологічне насичення спостерігається для впорядкованої множини, зібраної розробленим у роботі методом, і впорядкованої множини, вибраної з «Microsoft Academic» за автоматично призначеною категорією, для публікацій, знайдених в «ACM digital library» та «Google Академії», насичення не спостерігається.

Таблиця 2 показує, що вдосконалений у роботі метод бібліографічного виявлення та відбору дозволяє створити меншу мінімальну термінологічно насичену впорядковану множину публікацій, ніж розглянуті аналоги.

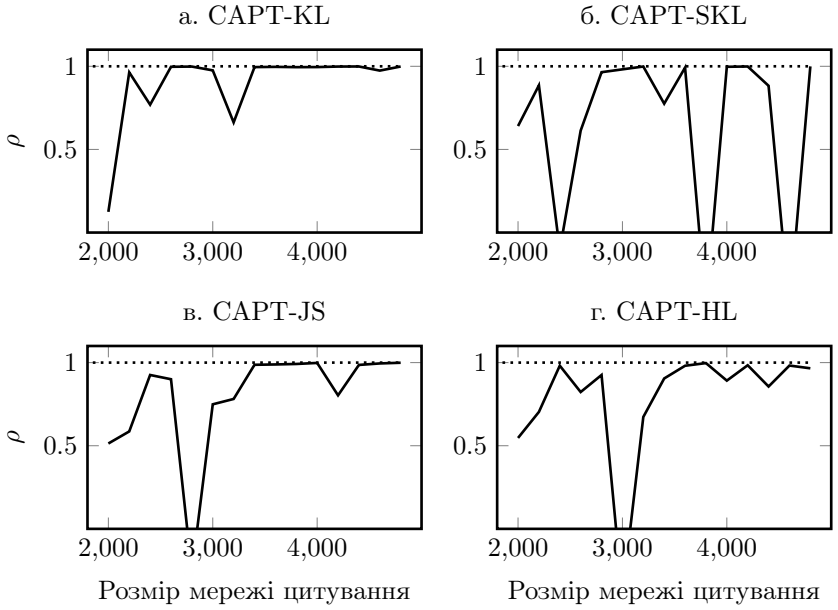

Рис. 2 Рангова кореляція Спірмена ρ між сусідніми членами послідовності впорядкованих множин публікацій, отриманих із поступовим збільшенням мереж цитування CAPT-KL, CAPT-SKL, CAPT-JS, CAPT-HL. Умова нерухомої точки $\rho = 1$ найточніше виконується для мережі CAPT-KL

Таблиця 2

Розмір мінімальних термінологічно насичених впорядкованих множин публікацій з теми «ontologies», отриманих різними методами.

Джерело	**Метод пошуку**	**Розмір**
«Microsoft Academic»	«Снігова куля»	160
«Microsoft Academic»	Автоматична мітка	180
«ACM digital library»	Авторська мітка	> 200
«Google Академія»	Ключові слова	≥ 220

Наводиться приклад побудови та аналізу мережі цитування для підготовки бібліографічного покажчика з методів персоналізації систем

комп'ютерного навчання вимові.

У додатках до дисертації наведені документи про використання отриманих результатів, звіти про практичну апробацію – приклади вихідних даних, створених на їх основі бібліографічних покажчиків, супутніх множин термінів та їх використання у процесі виконання кваліфікаційних, дипломних та наукових робіт.

ВИСНОВКИ

У дисертаційній роботі поставлене та вирішене актуальне науково-практичне завдання підвищення ефективності процесу бібліографічного виявлення й відбору шляхом розробки удосконалених моделі, методу та інформаційної технології шляхом розробки гібридної математичної моделі, методу бібліографічного виявлення та відбору та відповідних програмних рішень, що дозволяє підвищити ефективність процесу створення науково-допоміжного або рекомендаційного бібліографічного покажчика за рахунок зменшення кількості публікацій, які пропонується вивчити користувачу покажчика, при одночасному виконанні умови термінологічного насичення.

В процесі виконання роботи отримані такі наукові та практичні результати:

– Проаналізовано існуючі методи бібліографічного виявлення та відбору в наукометричних базах даних і з'ясовано, що якісні виявлення та відбір потребують вивчення експертами всіх виявлених публікацій, як релевантних, так і нерелевантних, із метою покращення пошукового запиту і досягнення більшої точності та повноти, що, в свою чергу, разом із відсутністю критеріїв повноти збільшує витрати часу.

– Виявлено моделі та методи, що можуть бути вдосконалені, об'єднані та застосовані для підвищення ефективності бібліографічного виявлення та відбору, а саме: ймовірнісне тематичне моделювання текстових документів, створення й аналіз мереж цитування, автоматичне виявлення термінів.

– Вперше на основі теорії множин, теорії графів, математичної статистики, методів обробки природних мов розроблено гібридну математичну модель процесу бібліографічного виявлення та відбору яка відрізняється від наявних одночасним використанням ітерацій контрольованої «снігової кулі», ймовірнісного тематичного моделювання текстових документів, аналізу мережі цитування, автоматичного виявлення термінів, наявністю ознак нерухомої точки ітерацій та показників насиченості набору термінів обраної предметної області, що

створило підґрунтя для розробки методу виявлення та відбору мінімальної термінологічно насиченої впорядкованої множини публікацій.

– Вперше на основі розробленої моделі розроблено метод бібліографічного виявлення та відбору, який відрізняється від наявних аналогів спільним узгодженим застосуванням вдосконалених методів відбору початкової множини публікацій, побудови ймовірнісної тематичної моделі текстових документів, контрольованої «снігової кулі», аналізу мережі цитування, автоматичного виявлення термінів та виявлення термінологічного насичення. На відміну від відомих аналогів, розроблений метод дозволяє у процесі бібліографічного виявлення та відбору швидше отримати мінімальну термінологічно насичену впорядковану множину публікацій.

– Вдосконалена імовірнісна тематична модель текстових документів шляхом застосування методу головних компонент, що дозволяє автоматично визначити кількість тем.

– На основі розробленої моделі та методу бібліографічного виявлення та відбору розроблено інформаційну технологію, що, на відміну від відомих аналогів, дозволяє створювати мінімальні термінологічно насичені впорядковані множини публікацій, формуючи коротші бібліографічні покажчики, які містять найважливіші терміни предметної області.

– Програмно реалізовано розроблену інформаційну технологію.

– Проведено експериментальне дослідження розробленої технології та для колекції ONTO-KL показано, що термінологічне насичення для бібліографічного покажчика, зібраного в базі «Microsoft Academic» розробленим методом бібліографічного виявлення та відбору, настає для 160 публікацій – на 9% швидше, ніж для вибірки з бази «Microsoft Academic» за автоматично призначеною категорією «ontology» (180 публікацій). Для перших 200 публікацій, знайдених за ключовим словом «ontology» в «Google Академії», та перших 200 публікацій, знайдених у «ACM digital library» за авторською міткою «ontology», насичення не спостерігається.

– Проведено практичну апробацію розробленої інформаційної технології в навчальному процесі Українського Католицького Університету та Запорізького національного університету.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні наукові результати дисертації:

Статті у наукових фахових виданнях України:

1. О. О. Тодоріко та Г. А. Добровольський, "Використання набору сигнатур n-грам для пошуку за схожістю," *Вісник Херсонського національного технічного університету*, №. 2(47), с. 343 – 347, 2013. *Здобувач брав участь у розробці методів пошуку коротких текстів за схожістю, плануванні та виконанні обчислювальних експериментів.*

2. Г. А. Добровольський, Н. Г. Кеберле та П. П. Прохоренко, "Застосування методів побудови та аналізу мережі цитування для підготовки бібліографічного покажчика з методів персоналізації систем комп'ютерного навчання вимові," *Вісник Херсонського національного технічного університету*, №. 3(66), с. 278 – 285, 2018. *Здобувачем розроблений метод виявлення та відбору застосовано для збору бібліографічного покажчика.*

3. Г. А. Добровольський та Н. Г. Кеберле, "Математична модель відбору наукових публікацій у процесі підготовки бібліографічного покажчика," *Вісник Кременчуцького національного університету імені Михайла Остроградського*, №. 1(120), с. 86 – 92, 2020. *Здобувачем розроблено модель виявлення та відбору наукових публікацій у процесі підготовки бібліографічного покажчика.*

Наукові праці, в яких опубліковані основні наукові результати дисертації у зарубіжних спеціалізованих виданнях:

4. H. Dobrovolskyi and N. Keberle, "Principal component analysis in topic modelling of short text document collections," in *CEUR Workshop Proceedings*, vol. 1851, pp. 48–54, CEUR-WS, 2017. **Scopus.** *Здобувач брав участь у розробці шляхів застосування методу головних компонент до тематичного моделювання текстових документів для порівняння бібліографічних анотацій.*

5. H. Dobrovolskyi and N. Keberle, "Collecting the seminal scientific abstracts with topic modelling, snowball sampling and citation analysis," in *CEUR Workshop Proceedings*, vol. 2105, pp. 179–192, CEUR-WS, 2018. **Scopus.** *Здобувачем розроблено інформаційну технологію бібліографічного виявлення та відбору.*

6. H. Dobrovolskyi and N. Keberle, "On convergence of controlled snowball sampling for scientific abstracts collection," in *Communications in Computer and Information Science*, vol. 1007, pp. 18–42, Springer Nature,

2019. **Scopus.** *Здобувачем запропоновано ознаки збіжності ітерацій контрольованої «снігової кулі», виконано обчислювальні експерименти.*

7. V. Kosa, D. Chaves-Fraga, H. Dobrovolskiy, E. Fedorenko, and V. Ermolayev, "Optimizing automated term extraction for terminological saturation measurement," in *CEUR Workshop Proceedings*, vol. 2387, pp. 1–16, CEUR-WS, 2019. **Scopus.** *Здобувач брав участь у розробці оптимізованого методу автоматичного виявлення термінів.*

8. V. Kosa, D. Chaves-Fraga, H. Dobrovolskyi, and V. Ermolayev, "Optimized term extraction method based on computing merged partial c-values," in *Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2019* (V. Ermolayev, F. Mallet, V. Yakovyna, H. Mayr, and A. Spivakovsky, eds.), vol. 1175 of *Communications in Computer and Information Science*, pp. 24–49, Springer Berlin Heidelberg, 2020. **Scopus.** *Здобувач брав участь у вдосконаленні оптимізованого методу автоматичного виявлення термінів.*

Наукові праці, які засвідчують апробацію матеріалів дисертації:

9. О. О. Тодоріко, Г. А. Добровольський та М. Г. Добровольська, "Застосування нейронної мережі для автоматичної класифікації коротких текстових документів," *Вісник Херсонського національного технічного університету*, №. 2(35), с. 421 – 425, 2009. *Здобувач брав участь у розробці методу порівняння і відбору бібліографічних анотацій. Форма участі: очна, доповідь на конференції.*

10. О. О. Тодоріко та Г. А. Добровольський, "Словниковий пошук за схожістю за допомогою хешів на основі сигнатур," *Вісник Херсонського національного технічного університету*, №. 3(39), с. 467 – 471, 2010. *Здобувач брав участь у розробці методів пошуку коротких текстів за схожістю, плануванні та виконанні обчислювальних експериментів. Форма участі: очна, доповідь на конференції.*

11. О. О. Тодоріко и Г. А. Добровольський, "Использование хеширования по нескольким сигнатурам для очистки и объединения словарей данных на примере названий географических объектов," *Вісник Херсонського національного технічного університету*, №. 3(42), с. 419 – 423, 2011. *Здобувач брав участь у розробці методів хешування, плануванні та виконанні обчислювальних експериментів з оцінки якості методів порівняння та відбору коротких текстів. Форма участі: очна, доповідь на конференції.*

12. О. А. Тодорико и Г. А. Добровольский, "Оценка сигнатурных алгоритмов поиска по сходству в словаре," *Вісник Херсонського національного технічного університету*, №. 2(41), с. 250 – 254, 2011.

Здобувач брав участь у плануванні та виконанні обчислювальних експериментів з оцінки якості методів порівняння та відбору коротких текстів. Форма участі: очна, доповідь на конференції.

13. О. О. Тодоріко та Г. А. Добровольський, "Оцінка імовірності колізій для хеш-функцій, які представляють слово за допомогою набору сигнатур," *Вісник Херсонського національного технічного університету*, №. 2(45), с. 383 – 388, 2012. *Здобувач брав участь у розробці математичної моделі пошуку за схожістю, плануванні та виконанні експериментів з оцінки якості хеш-функцій, призначених для порівняння та відбору коротких текстів. Форма участі: очна, доповідь на конференції.*

14. О. О. Тодоріко та Г. А. Добровольський, "Програмний інструментарій для пошуку за схожістю та зіставлення записів," *Вісник Херсонського національного технічного університету*, №. 1(44), с. 204 – 208, 2012. *Здобувач брав участь у розробці структури та інтерфейсів бібліотеки класів, призначеної для пошуку за схожістю та зіставлення структур даних. Форма участі: очна, доповідь на конференції.*

15. M. Davidovsky, G. Dobrovolsky, O. Todoriko, and V. Davidovsky, "Adaptable enterprise information systems development using advanced active data dictionary framework," in *Information Systems: Methods, Models, and Applications* (H. Mayr, C. Kop, S. Liddle, and A. Ginige, eds.), vol. 137 of *Lecture Notes in Business Information Processing*, pp. 152–161, Springer Berlin Heidelberg, 2013. **Scopus.** *Здобувач брав участь у розробці інформаційної технології на основі потоків даних. Форма участі: очна, доповідь на конференції.*

16. H. Dobrovolskyi, N. Keberle, and Y. Ternovyi, "Sparse symmetric nonnegative matrix factorization applied to face recognition," in *Proceedings of the 2017 IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2017*, vol. 2, pp. 1042–1045, Springer, 2017. **Scopus.** *Здобувач брав участь у розробці вдосконаленого методу розрідженої симетричної невід'ємної матричної факторизації. Форма участі: очна, доповідь на конференції.*

17. H. Dobrovolskyi, N. Keberle, and O. Todoriko, "Probabilistic topic modelling for controlled snowball sampling in citation network collection," in *International Conference on Knowledge Engineering and the Semantic Web*, pp. 85–100, Springer, 2017. **Scopus.** *Здобувачем розроблено метод застосування ймовірнісного тематичного моделювання текстових документів для контролю розмірів «снігової кулі». Форма участі: очна,*

доповідь на конференції.

18. М. Чміль та Г. А. Добровольський, "Міжмодульний механізм зв'язку для створення масштабованої інформаційної системи підприємства," в *Збірка тез доповідей Восьмої Всеукраїнської, п'ятнадцятої регіональної наукової конференції молодих дослідників «Актуальні проблеми Математики та інформатики»*. Запоріжжя: ЗНУ, 2017. с. 65. *Форма участі: очна, доповідь на конференції.*

19. Г. А. Добровольський та К. М. Кучерян, "Реалізація алгоритму підрахунку шляхів пошуку для аналізу мережі цитування," в *Збірка тез доповідей Дев'ятої Всеукраїнської, шістнадцятої регіональної наукової конференції молодих дослідників «Актуальні проблеми Математики та інформатики»*. Запоріжжя: ЗНУ, 2018. с. 46. *Форма участі: очна, доповідь на конференції.*

Наукові праці, які додатково відображають наукові результати дисертації:

20. О. О. Тодоріко та Г. А. Добровольський, "Оцінка якості автоматичної класифікації коротких текстових документів," *Вісник Запорізького національного університету : зб. наук. пр. Фізико-математичні науки.*, №. 2, с. 131 – 140, 2010. *Здобувач брав участь у плануванні та виконанні обчислювальних експериментів із оцінки якості порівняння та відбору бібліографічних анотацій.*

21. О. О. Тодоріко та Г. А. Добровольський, "Спосіб пошуку текстової інформації за схожістю," Патент України № 71159, 10.07.2012. *Здобувач брав участь у розробці способу пошуку за схожістю.*

АНОТАЦІЯ

Добровольський Г. А. Модель, метод та інформаційна технологія відбору наукових публікацій у процесі підготовки бібліографічного покажчика. — Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 — інформаційні технології. — Запорізький національний університет Міністерства освіти і науки України; Харківський національний університет імені В. Н. Каразіна Міністерства освіти і науки України. — Харків, 2021.

У дисертаційній роботі поставлена та вирішена актуальна науково-практична задача удосконалення засобів бібліографічного виявлення та відбору у процесі створення науково-допоміжного або рекомендаційного бібліографічного покажчика. Розроблено гібридну математичну модель процесу, яка поєднує ітерації контрольованої «снігової кулі», вдосконалену ймовірнісну тематичну модель текстових документів, аналіз мережі цитування, автоматичне виявлення термінів, ознаки нерухомої точки ітерацій та насиченості набору термінів. Розроблено вдосконалений ітеративний метод бібліографічного виявлення та відбору, що пропонує шлях практичного отримання описаної моделлю мінімальної термінологічно насиченої впорядкованої множини документів. Експериментальне дослідження показало, що термінологічно насичений бібліографічний покажчик, зібраний в базі «Microsoft Academic» розробленим методом, містить на 9% менше публікацій, ніж вибірка за автоматично призначеною категорією. На основі розроблених моделі та методу розроблено та програмно реалізовано інформаційну технологію, проведено її практичну апробацію в навчальному процесі Українського Католицького Університету та Запорізького національного університету.

Ключові слова: автоматичне виявлення термінів, аналіз головних шляхів, бібліографічне виявлення, бібліографічний відбір, імовірнісна тематична модель текстових документів, інформаційна технологія, математична модель, мережа цитування, метод головних компонент, метод контрольованої снігової кулі, метод виявлення та відбору, низька специфічність інформаційної потреби, розріджена симетрична невід'ємна матрична факторизація, термінологічне насичення.

ABSTRACT

Dobrovolskyi H. A. Model, Method and Information Technology of Scientific Publication Selection in the Process of Bibliographical Index Compilation. — Qualification scientific work is as a manuscript.

Thesis for a Candidate Degree in Technical Sciences, Speciality 05.13.06 — Information technologies. — Zaporizhzhia National University of Ministry of Education and Science of Ukraine; V. N. Karazin Kharkiv National University of Ministry of Education and Science of Ukraine. — Kharkiv. 2021.

The dissertation presents a model, a method and an information technology for automatic gathering and updating of the personalized scientific-supporting or recommender bibliographical indexes. The objective of the research is to increase the efficiency of the gathering process providing the indexes containing all significant publications in the domain of interest. At the same time the indexes are small enough to be studied by a human in an acceptable time.

In the dissertation, the discovery and selection method is developed as the snowball sampling combined with probabilistic topic model of the publication texts, citation analysis of the collected citation network, automatic term extraction, and terminological saturation detection. The bibliographic index built in such a way appears to be the minimal terminologically saturated ordered publication set that can be read by an expert in minimal time. The corresponding criterion of efficiency is the number of publications in the index.

The scientific novelty of the obtained results is:

– For the first time, the hybrid mathematical model of the process of bibliographic discovery and selection is developed. The distinctive feature of the model is the combination of snowball iterations, probabilistic topic model of the text documents, citation network analysis, automatic term extraction, the condition of the iteration stationary point, and the condition of terminological saturation in the domain of interest.

– For the first time, the method of bibliographic discovery and selection is developed. The method consists of the coordinated application of the improved methods of seed collection picking, probabilistic topic model construction, controlled snowball iterations, citation network analysis, automatic term extraction and terminological saturation detection.

– For the first time the information technology is developed to implement the model and the method.

Unlike analogs, the model, the method, and the technology allow the creation of the minimal terminologically saturated bibliographic indexes that

contain the seminal terms of the scientific domain of interest.

– The method of probabilistic topic modeling of the text documents is improved with the application of the principal component approximation. The improvement allows determining the number of topics in an automatic way.

– The information retrieval methods are advanced in the field of meeting the low-specific information need. The essence of the advance is the stopping search criterion that allows minimizing the search time with collecting the minimal terminologically saturated ordered publication.

The key experimental study looks into terminological saturation of the ordered publication set that was collected with the developed controlled snowball sampling method. For the chosen domain of interest "Ontologies (computer science)", it is shown that the terminological saturation is observed if the collection is gathered with the developed controlled snowball sampling method in "Microsoft Academic" database and if the collection is gathered with automatically assigned keyword "Ontologies (computer science)". For the publications found with a keyword search in the "Google Scholar" service and the publications found in ACM digital library terminological saturation is not observed and, perhaps, requires collecting the larger publication set. The calculated quality measure – the size of minimal terminologically saturated publication set shows that the developed information technology allows the 9% less minimal terminologically saturated publication set than the best of the considered analogs.

The developed and implemented information technology was successfully applied in the learning process of the Faculty of Applied Sciences of Ukrainian Catholic University as part of course "Automated Term Extraction and Ontology Learning from Texts" that is included in the Data Science Master's degree program. Another use case of the developed technology is its application in the learning process of the Computer Science department of Zaporizhzhia National University to collect the seminal publications and then write the Related Work section of the bachelor's and master's thesis.

Key words: automatic term extraction, bibliographic disovery, bibliographic selection, citation network, controlled snowball sampling, information technology, low-specific information need, main path analysis, mathematical model, method of disovery and selection, principal component analysis, probabilistic topic modelling of text documents, sparse symmetric nonnegative matrix factorization, terminological saturation.

Підписано до друку 03.02.2021. Формат 60 × 84/16. Папір друк. Офсет. друк. Фіз. друк. арк. 1,5. Умовн. друк. арк. 1,4. Тираж 100 пр. Зам. 89.

Дані про друкарню