УДК 004.8:004.912:001.811 DOI: 10.30929/1995-0519.2020.1.86-92

МАТЕМАТИЧНА МОДЕЛЬ ВІДБОРУ НАУКОВИХ ПУБЛІКАЦІЙ У ПРОЦЕСІ ПІДГОТОВКИ БІБЛІОГРАФІЧНОГО ПОКАЖЧИКА

Г. А. Добровольський, Н. Г. Кеберле
Запорізький національний університет
вул. Жуковського, 66, м. Запоріжжя, 69600, Україна. E-mail: gen.dobr@gmail.com

Представлено гібридну математичну модель процесу бібліографічного виявлення та відбору наукових публікацій, що описує процедуру, спрямовану на задоволення інформаційної потреби у рекомендаційному або науково-допоміжному бібліографічному покажчику, у якому всі документи належать до вибраної користувачем предметної області, мають у ній важливе значення, описують її основні твердження, можуть бути вивчені за прийнятний користувачем час та додані як перелік посилань у наукову публікацію обмеженого обсягу. Модель було сформульовано з метою послідовного формального опису розробленого авторами методу контрольованої «снігової кулі» у термінах теорії графів, теорії ймовірностей та теорії множин. Основними компонентами представленої моделі є ймовірнісна тематична модель, відображення ітерацій контрольованої «снігової кулі», умови нерухомої точки для ітерацій, аналіз шляхів у мережі цитування, автоматичне виявлення термінів. Модель відрізняється від наявних аналогів поєднанням названих компонент, застосуванням симетричної розрідженої невід'ємної матричної факторизації для тематичного моделювання, а також наявністю ознак нерухомої точки ітерацій та показників насиченості набору термінів обраної предметної області, що створило підґрунтя для реалізації методу бібліографічного виявлення та відбору і відповідної інформаційної технології. Результатом виконаної роботи стали: визначення рекомендаційного та науково-допоміжного бібліографічного покажчика, як мінімальної термінологічно насиченої впорядкованої множини документів; визначення міри якості моделі, як розміру згаданої множини. Також запропоновано спрощені критерії збіжності ітеративного процесу бібліографічного виявлення та відбору.

Ключові слова: бібліографія, відбір, мережа цитування, тематична модель, виявлення термінів, міра якості, термінологічне насичення.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ОТБОРА НАУЧНЫХ ПУБЛИКАЦИЙ В ПРОЦЕССЕ ПОДГОТОВКИ БИБЛИОГРАФИЧЕСКОГО УКАЗАТЕЛЯ

Г. А. Добровольский, Н. Г. Кеберле
Запорожский национальный университет
ул. Жуковского, 66, г. Запорожье, 69600, Украина. E-mail: gen.dobr@gmail.com

Представлена гибридная математическая модель процесса библиографического выявления и отбора научных публикаций, которая описывает процедуру, направленную на удовлетворение информационной потребности в рекомендательном или научно-вспомогательном библиографическом указателе, в котором все документы относятся к выбранной пользователем предметной области, имеют в ней важное значение описывают ее основные утверждения, могут быть изучены за приемлемое пользователем время и добавлены в качестве перечня ссылок к научной публикации ограниченного объема. Модель была сформулирована с целью последовательного формального описания разработанного авторами метода контролируемого «снежного шара» в терминах теории графов, теории вероятностей и теории множеств. Основными компонентами представленной модели является вероятностная тематическая модель, отображение итераций контролируемого «снежного шара», условие неподвижной точки для итераций, анализ путей в сети цитирования, автоматическое обнаружение терминов. Модель отличается от имеющихся аналогов сочетанием названных компонент, применением симметричной разреженной неотрицательной матричной факторизации для тематического моделирования, а также наличием признаков неподвижной точки итераций и показателей насыщенности набора терминов выбранной предметной области, создавая основу для реализации метода библиографического выявления и отбора и соответствующей информационной технологии. Результатом проведенной работы стали: определение рекомендательного и научно-вспомогательного библиографического указателя, как минимального терминологически насыщенного упорядоченного множества документов; определение показателя качества модели, как размера упомянутого множества. Также предложены упрощенные критерии сходимости итеративного процесса библиографического выявления и отбора.

Ключевые слова: библиография, отбор, сеть цитирования, тематическая модель, вывление терминов, мера качества, терминологическое насыщение.

АКТУАЛЬНІСТЬ РОБОТИ. Огляд літератури є важливою компонентою кожного дослідницького проекту. Він є основою для створення нових знань, спрощує розвиток теорії, завершує дослідження у відомих областях знання та відкриває нові області досліджень. Першим кроком до якісного огляду є пошук публікацій, що містять основні наукові результати – складання науково-допоміжного або рекомендаційного бібліографічного покажчика [1]. Звичайно, створення покажчика вимагає достатніх знань з теми дослідження та складається із [2, 3] формулювання актуальної мети огляду, систематичного виявлення та відбору релевантних публікацій, аналізу та викладення результатів огляду у вигляді статті. Процес формування огляду, який керується згаданими принципами, завжди дає задовільний результат, але є трудомістким, тому його автоматизація є актуальною задачею.

Виявлення та відбір публікацій найчастіше виконуються [1, 2, 4, 5] в електронних базах даних шляхом пошуку за ключовими словами. Недоліком пошуку за ключовими словами в електронних базах є упередженість — систематична помилка при складанні пошукового запиту, яка з'являється з різних причин [3]. Щоб подолати упередженість та скласти коректний набір ключових слів, Petticrew та Gilbody [3] радять провести опитування експертів. У випадку недоступності експертів корекція набору ключових слів [4, 5] виконується шляхом вивчення всіх виявлених документів – як релевантних, так і нерелевантних, що спричиняє зайві витрати часу і не гарантує повноти корпусу. Необхідність корекції набору ключових слів також виявляється або за допомогою аналізу відібраних документів або за допомогою опитування експертів.

Виявлення також може виконуватися методом «снігової кулі» [6, 7, 8], коли кожна зі знайдених публікацій досліджується і до результатів пошуку включаються усі статті, що її цитують, та роботи, на які вона посилається. Метод «снігової кулі» здатен знайти документи, які містять ключові слова, про які користувач не знає, і з великою ймовірністю виявляє документи, які часто цитуються.

Через те, що сучасні системи пошуку наукових публікацій містять велику кількість записів, важливою частиною стратегії пошуку є критерій завершення виявлення та відбору. Для великої кількості виявлених документів пропонується [3] зупинитися в момент насичення множини публікацій, коли дослідник зрозуміє, що врахування в огляді нових публікацій не впливає на зроблені висновки. Виявити насичення допомагає виявлення та оцінка важливості концепцій вибраної області досліджень [2]. Однак згадані методики [2, 3] не пропонують формальних критеріїв зупинки процесу виявлення та відбору.

Метою даної роботи є презентація формальної гібридної математичної моделі процесу бібліографічного виявлення та відбору, що є підґрунтям для розробленого авторами методу бібліографічного виявлення та відбору важливих публікацій [8], яка відрізняється від найближчих аналогів [6, 9, 10] одночасним використанням ітеративного методу контрольованої «снігової кулі», ймовірнісного тематичного моделювання текстових документів, аналізу мережі цитування, автоматичного виявлення термінів та наявністю формальних умов насичення множини публікацій.

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ.

Означення.

Процес бібліографічного виявлення та відбору – це процес вибору із множини всіх доступних для аналізу документів D деякої підмножини $B \subseteq D$, елементи якої задовольняють інформаційну потребу користувача бібліографічного покажчика.

Термін «*слово»* позначає одиницю мови, яка виражає окреме поняття.

Словосполученням c називається слово або кортеж слів.

Речення s – граматична конструкція, побудована з одного чи кількох словосполучень, яка виражає окрему, відносно незалежну думку.

Терміном τ є словосполучення, що позначає поняття заданої предметної області.

Документом або *публікацією* в роботі називається структура із кількох текстів, термінів, метаінформації та посилань на інші публікації. Типовими частинами публікації є її унікальний ідентифікатор (наприклад, DOI), назва, анотація, повний текст, ключові слова (множина термінів). Метаінформація найчастіше містить імена авторів, посилання на елементи класифікатора. Посиланнями на інші публікації є множина їх унікальних ідентифікаторів. Таблиці, формули, ілюстрації, в даній роботі не розглядаються.

Відображення цитування на множині документів D визначається як

$$R: v \to \{u \in D \mid v \text{ цитує } u, v \in D\}, \qquad (1)$$

застосувавши відображення цитування до деякого документа u, можна отримати множину документів, на які він посилається. Обернене відображення цитування визначається як

$$R^{-1}: u \to \{v \in D \mid v \text{ цитує } u, u \in D\}. \qquad (2)$$

Застосувавши обернене відображення цитування до деякого документа u, можна отримати множину документів, які посилаються на нього. Повторне відображення цитування $R^k, k \in \mathbb{N}$ визначається, як результат багаторазового застосування операції R:

$$R^k = \begin{cases} R, k = 1 \\ R \cdot R^{k-1}, k > 1 \end{cases}. \qquad (3)$$

Відображення (1) визначає орієнтований граф – *мережу цитування* $N = (D, E)$ з дугами $E = \{vu, \forall u \in D, v \in D, u \in R(v)\}$ та вершинами $d \in D$.

Інформаційна потреба – це неформальна і часто неявно виражена сукупність вимог до результатів інформаційного пошуку [11]. Документ є *релевантним*, якщо він з точки зору користувача задовольняє інформаційну потребу.

Припущення.

Розроблена математична модель процесу бібліографічного виявлення та відбору спирається на такі припущення:

Припущення 1. Інформаційна потреба складається із кількох неформальних вимог:

1) належності всіх документів із B до заданої предметної області;

2) важливості всіх документів із B у заданій предметній області;

3) таку кількість всіх документів у B, що дозволяє їх детальне вивчення користувачем за прийнятний час та укладання переліку посилань у науковій публікації обмеженого об'єму;

4) наявність у документах B усіх основних термінів предметної області.

Надалі вважається, що інформаційна потреба частково виражається користувачем у формі множини документів, кожен із яких стосується обраної користувачем теми. При цьому користувач знає деякі

ключові слова із заданої області знань і може відібрати документи потрібної тематики, але не має достатньої кваліфікації для оцінки їх важливості та повноти зібраного бібліографічного покажчика.

Припущення 2. Кожен документ $d \in D$ можна відобразити у множину речень $S(d)$, і кожне речення $s \in S(d)$ – у множину словосполучень $C(s)$, яка є підмножиною всіх словосполучень C, які зустрічаються у D.

Припущення 3. На множині доступних для аналізу документів D визначене відображення цитування.

Припущення 4. Мережа цитування N є майже ациклічною [12]

$$|\{d \in D \mid \exists\, k \in \mathbb{N}, d \in R^k(d)\}| \ll |\{d \in D \mid \forall\, k \in \mathbb{N}, d \notin R^k(d)\}|, \quad (4)$$

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

Припущення 5. Необхідною ознакою наявності у В всіх основних термінів предметної області є термінологічне насичення впорядкованої множини документів [13].

Припущення 6. Повний текст публікації недоступний. Часто через обмеження, встановлені власниками авторських прав, доступ до повного тексту публікації ускладнений.

Математична модель.

Розроблена формальна гібридна математична модель процесу бібліографічного виявлення та відбору – це кортеж

$$\mathcal{M} = \langle D, R, B_0, PTM, diff, diff_{max}, \hat{B}, SPC, \Delta, \Delta_{min}, M, \hat{T}, Cvalue, thd \rangle, \quad (5)$$

де D – доступні для аналізу документи; R – відображення цитування; B_0 – початкова точка ітерацій «снігової кулі»; PTM – представлення змісту документа; $diff$ – міра відмінності документів; $diff_{max}$ – гранична міра відмінності документів; $\hat{B}$ – відображення ітерацій контрольованої «снігової кулі»; SPC – вага документа у предметній області; Δ – міра близькості впорядкованих множин документів; Δ_{min} – границя міри близькості впорядкованих множин документів; M – максимальний ранг документів; $\hat{T}$ – відображення документів із D у множину термінів T; $Cvalue(\tau)$ – вага терміна τ; thd – міра відмінності множин термінів.

Початковою точкою ітерацій «снігової кулі» є множина релевантних документів B_0 ($B_0 \subseteq D$, $|B_0| \sim O(10)$), вибрана користувачем, наприклад, шляхом пошуку за ключовими словами. Вона є формою часткового вираження інформаційної потреби – належності всіх документів до заданої предметної області. Множина B_0 разом із представленням змісту документа PTM, мірою відмінності документів $diff$ та граничною мірою відмінності $diff_{max}$ дозволяє визначити належність довільного виявленого документа до предметної області.

Варто відмітити, що неповнота знань користувача про предметну область може стати причиною деформації початкової множини B_0, тому особливо важливо дослідити залежність результату ітерацій контрольованої «снігової кулі» від варіації B_0.

Представлення змісту документа знаходиться за допомогою імовірнісної тематичної моделі (ІТМ), яка відображає зміст кожного документа $d \in D$ у кортеж умовних ймовірностей $p(t|d) = PTM(d)$, що показують імовірність належності документа d до теми t. Кожна тема t визначається ймовірностями $p(C_i|\, t)$ належності кожного словосполучення C_i до теми t, і апріорною ймовірністю $p(t)$. Назви та анотації документів є короткими текстами, тому найефективнішим є використання у моделі (5) ІТМ на основі парних ймовірностей слів [8, 14], яка будується шляхом визначення розподілів $p(c_i|\, t)$ та $p(t)$, виходячи з частоти сусідства словосполучень

$$p(c_i, c_k) = \sum_t p(c_i|t)p(t)p(c_k|t), \quad (6)$$

яка підраховується як кількість речень, у яких одночасно зустрілися c_i та c_k. Використана ІТМ відрізняється від найближчих аналогів [15, 16], використанням методу головних компонент на основі симетричної розрідженої невід'ємної матричної факторизації, що дозволяє знайти необхідну для побудови моделі кількість тем [8, 14].

Представлення документів у вигляді набору ймовірностей дозволяє використати у якості міри відмінності документів $diff$ найпоширеніші способи порівняння тематичних моделей – дивергенцію Кульбака-Лейблера, симетричну дивергенцію Кульбака-Лейблера, відстань Гелінгера та дивергенцію Дженсена-Шеннона [17]. Граничне значення міри відмінності $diff_{max}$ є вхідним параметром моделі, який керує швидкістю росту «снігової кулі». У запропонованій моделі міра відмінності одного документа та початкової множини документів обчислюється за допомогою застосування ймовірнісної тематичної моделі до одного з найпростіших варіантів міжкластерної відстані – відстані до найближчого сусіда [18]

$$diff(v, B_0) = min[diff(u, v)|u \in B_0]. \quad (7)$$

На відміну від мір, що базуються на центроїдах, міра (7) не вимагає певної форми від кластерів, що послаблює вимоги до початкової множини документів. Умова

$$diff(v, B_0) < diff_{max}\ , \quad (8)$$

є способом виконання першої вимоги інформаційної потреби у рамках розробленої моделі.

Для визначення ітеративного методу контрольованої «снігової кулі» розглянемо максимальну мережу цитування $N_\infty = (D_\infty, E_\infty)$, яка містить всі доступні наукові публікації. Відповідно, задачу про бібліографічне виявлення і відбір можна інтерпретувати як задачу про знаходження деякої підмножини вузлів повної мережі цитування. Одним із варіантів її вирішення [6, 9, 10, 14] є алгоритм «снігової кулі». Основою алгоритму є наступні ітерації: із множини B_i вибирається публікація v, і всі публікації $R(v)$, на які вона посилається, додаються до множини B_{i+1}. Інший варіант ітерацій використовує зворотне відображення цитування R^{-1}, формуючи множину публікацій, які посилаються на v. Також можливо скомбінувати обидва варіанти ітерацій.

Недоліком ітерацій методу «снігової кулі» є значне зростання розміру множини B_{i+1} з ростом номера ітерації i. Тому необхідним компонентом є метод контролю кількості елементів $|B_{i+1}|$ – розміру «снігової кулі», використовуючи спостереження, що майже кожен список посилань містить документи, безпосередньо не пов'язані з досліджуваною темою. У даній роботі пропонується застосування міри схожості, обчисленої методом імовірнісного тематичного моделювання. Перевагою тематичної моделі є можливість обчислення тематики документа будь-якої довжини, і врахування значень термінів.

Таким чином, використане в моделі відображення ітерацій «снігової кулі», призначене для бібліографічного виявлення, набуває вигляду:

$$B_{i+1} = \hat{B}(B_i) = \bigcup_{v \in B_i}\{v\} \cup R(v) \cup R^{-1}(v)\,|_{diff(v,B_0)<diff_{max}}\ , \quad (9)$$

де $B_i \subseteq D$ і $diff\ (v, B_0\ )$ – міра відмінності публікації v та початкової множини публікацій B_0, а $diff_{max}$ – максимальна міра відмінності. $diff_{max}$ є параметром алгоритму. Рівняння (9) відрізняється від аналогів [9, 10] використанням тематичної моделі для обчислення відмінності документів та переходами як за цитуваннями так і в протилежному напрямку.

Вага документа у предметній області $SPC_i: v \to \mathbb{N}, v \in D$ обчислюється на основі інформації про цитування і визначається шляхом аналізу головних шляхів пошуку(Search Path Count, SPC) [12] в підграфі N_i мережі цитування N Для цього у мережу додаються дві вершини – початкова b та кінцева e. Вершина, яка відповідає публікації, що не посилається на жодну релевантну публікацію, отримує вихідну дугу до вершини e, і вершина, яка відповідає публікації, на яку не посилається жодна релевантна публікація, отримує вхідну дугу від вершини b. Вагою $SPC(u, v)$ дуги (u, v), де $v \in R(u)$, вважається кількість різних шляхів від b до e, які проходять через дугу (u, v). Умовою успішного обчислення міри SPC_i є відсутність циклів у графі N_i, яка в реальних мережах цитування не гарантується. Тому для застосування міри SPC_i потрібно перетворити мережу цитування в ациклічний граф [13]. Міра SPC_i дозволяє знайти ранг $Rank_i(v)$ кожного документа у множині B_i:

$$Rank_i(v) = |\{u \in B_i |\, SPC_i(u) > SPC_i(v)\}| \quad (10)$$

та визначити шукану впорядковану множину документів:

$$L_i(M) = (v_k)_{k=1}^{|B_i|}, Rank_i(v_k) < M, \quad Rank_i(v_k) < Rank_i(v_{k+1}), \quad (11)$$

де максимальний ранг документів M обмежує кількість документів у бібліографічному покажчику і визначається вимогою досягнення нерухомої точки ітерацій (9) та наявністю термінологічного насичення. Рівняння (11) є виконанням вимоги щодо важливості всіх документів із $L_i(M)$ у заданій предметній області.

Термінологічне насичення впорядкованої множини документів визначається вимогою того, що додавання K документів у кінець впорядкованої множини (11) майже не змінює перелік термінів

$$thd\big(T_i(M), T_i(M+K)\big)/\epsilon_i < 1, \quad (12)$$

де відображення $\hat{T}$ документів $L_i(M)$ у множину термінів $T_i(M)$ відбувається шляхом застосування до об'єднаного тексту документів процедури автоматичного визначення термінів [13,19], яка визначає вагу $Cvalue_i(\tau)$ терміна τ у множині документів $L_i(M)$, граничне значення ϵ_i ваги терміна, та міру відмінності множин термінів $thd(T_i, T_j)$ [13].

Мінімальна термінологічно насичена впорядкована множина документів визначається рівнянням (11), у якому

$$M = min\left\{ X \left| \frac{thd(T_i(X), T_i(X+K))}{\epsilon_i} < 1 \right.\right\}, \quad (13)$$

де $K \in \mathbb{N}$. Умова (12) є виконанням вимог інформаційної потреби щодо кількості документів у результатах відбору, яка дозволяє детальне вивчення користувачем за прийнятний час та одночасної наявності у відібраних документах всіх термінів предметної області.

Мірою ефективності розробленої моделі (5) є кількість документів $|L_i(M)|$ у бібліографічному покажчику, визначеному умовами (11), (12) та (13).

Збіжність ітерацій.

Основою розробленої математичної моделі є ітераційна побудова впорядкованої множини документів, і, як для всіх ітераційних підходів, важливим є питання збіжності та умов зупинки ітерацій. Навіть якщо збіжність не доведена, то критерієм зупинки звичайно вважають досягнення стаціонарної точки, коли додаткові ітерації не змінюють результат.

Недоліком процесу оцінки термінологічного насичення є значні витрати часу на отримання із різних джерел повних текстів відібраних документів та їх подальшу обробку. Тому існує необхідність у використанні спрощеної умови збіжності, яка буде індикатором змін впорядкованої множини документів (11).

В рамках розробленої моделі мірою близькості впорядкованих множин документів Δ вибрана рангова кореляція Спірмена $\rho(L_i(M), L_{i+1}(M))$ і умовою нерухомої точки ітерацій (9) є нерівність

$$|\rho(L_i(M), L_{i+1}(M)) - 1| < \omega, i > i_0, \quad (14)$$

де ω – границя міри близькості впорядкованих множин документів (11), параметр моделі, який задає рівень варіативності впорядкованої множини документів.

Другою мірою змін була вибрана верхня границя довірчого інтервалу розподілу Пуассона для події $Y = v \in L_{i+1}(M) \wedge v \notin L_i(M)$, тобто імовірності того, що важливий документ v було знайдено тільки на $i+1$-му кроці ітерацій (9). Для оцінки довірчого інтервалу підраховувалась кількість позитивних рішень n_j для публікацій-кандидатів з порядковими номерами $j \le i \le j + Z$ та для рівня значимості $\alpha = 0{,}05$ і верхня границя довірчого інтервалу

порівнювалась із вибраною заздалегідь граничною ймовірністю $p_{top} = 0{,}05$:

$$\frac{1}{Z} G^{-1}\left(n_j + 1, 1 - \frac{\alpha}{2}\right) < p_{top} \quad , \quad (15)$$

де G^{-1} -- обернена регуляризована неповна нижня гамма функція, і $Z = 1000$.

Умови збіжності (14) та (15) мають на меті надання швидкої попередньої оцінки збіжності ітерацій, яка вимагає набагато менше часу, ніж визначене моделлю термінологічне насичення. Але для обґрунтованого застосування умов (14) та (15) потрібно детально дослідити їх зв'язок із умовою термінологічного насичення (12).

Для демонстрації збіжності було виконано ітерації (9) контрольованої «снігової кулі» для виявлення та відбору публікацій з теми «Методи комп'ютерного навчання вимові» [8]. Результати обчислень, показані на рис. 1, демонструють зменшення ймовірності успішного відбору значимої публікації зі збільшенням кількості виявлених публікацій. Ознакою насичення є зменшення ймовірності $P(Y)$ та обмежена кількість публікацій, які потрібно переглянути, щоб виконати умову $P(Y) < p_{top}$. Таким чином, можливість насичення множини відібраних публікацій зі збільшенням кількості виявлених публікацій отримує експериментальне підтвердження.

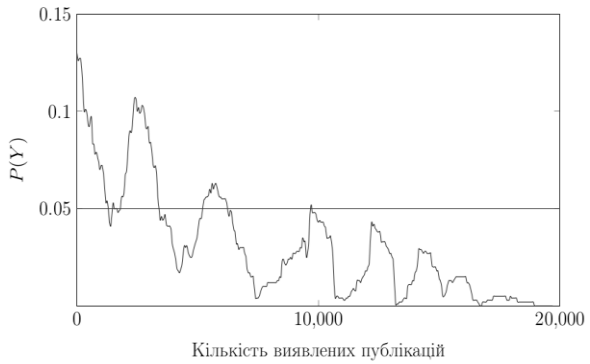

Рисунок 1 – Ймовірність P (Y) успішного відбору значимої публікації, як функція кількості виявлених публікацій

ВИСНОВКИ. У роботі представлено гібридну математичну модель процесу бібліографічного виявлення та відбору, що описує процедуру виявлення та відбору наукових публікацій, спрямовану на задоволення інформаційної потреби у рекомендаційному або науково-допоміжному бібліографічному покажчику, у якому всі документи належать до вибраної користувачем предметної області, мають у ній важливе значення, описують її основні положення, можуть бути вивчені за прийнятний користувачем час та додані до переліку посилань у наукову публікацію.

Модель було сформульовано з метою послідовного формального опису розробленого авторами методу контрольованої «снігової кулі» у термінах теорії графів, теорії ймовірностей та теорії множин.

Основними компонентами представленої моделі є ймовірнісна тематична модель, відображення ітерацій контрольованої «снігової кулі», умови нерухомої точки для ітерацій, аналіз мережі цитування, автоматичне виявлення термінів.

Модель відрізняється від наявних аналогів поєднанням названих компонент, застосуванням симетричної розрідженої невід'ємної матричної факторизації для тематичного моделювання, а також наявністю ознак нерухомої точки ітерацій та показників насиченості набору термінів обраної предметної області, що створило підґрунтя для реалізації методу бібліографічного виявлення та відбору і відповідної інформаційної технології.

Результатом виконаної роботи стали: визначення рекомендаційного та науково-допоміжного бібліографічного покажчика, як мінімальної термінологічно насиченої впорядкованої множини документів; визначення міри ефективності моделі, як розміру згаданої множини. Також запропоновано спрощені критерії збіжності ітеративного процесу бібліографічного виявлення та відбору.

Формалізація шляхом розробки математичної моделі розроблених раніше методу обчислень та інформаційної технології дозволила окреслити область їх застосування, ввести числові міри ефективності та збіжності, а також спланувати подальші експерименти з метою дослідження стійкості моделі та зв'язку спрощених критеріїв збіжності із термінологічним насиченням.

ЛІТЕРАТУРА

1. Галганова О. Складання бібліографічних посібників: інформ.-метод. матеріали. Київ: М-во культури України, Нац. парлам. бібліотека України, 2015.

2. Fisch C., Block J. Six tips for your (systematic) literature review in business and management research. *Management Review Quarterly*. 2018. Vol. 68, No. 2. Pp. 103–106.

3. Petticrew M., Gilbody S. Planning and conducting systematic reviews. *Health psychology in practice*. 2004. Pp. 150–179.

4. Friday D., Ryan S., Sridharan R., Collins D. Collaborative risk management: a systematic literature review. *International Journal of Physical Distribution & Logistics Management*. 2019. Vol. 48, No. 3. Pp. 231–253.

5. Colicchia C., Strozzi F. Supply chain risk management: a new methodology for a systematic literature review. *Supply Chain Management:An International Journal*. 2012. Vol. 17, No. 4. Pp. 403–418.

6. Mapping the historical development of physical activity and health research: A structured literature review and citation network analysis. / Varela A. et al. *Preventive Medicine*. 2018. Vol. 111. Pp. 466–472.

7. Liu J., Lu L., Lu W., Lin B. Data envelopment analysis 1978–2010: A citation-based literature survey. *Omega*, 2013. Vol. 41, No. 1. Pp. 3–15.

8. Dobrovolskyi H., Keberle N. Collecting the Seminal Scientific Abstracts with Topic Modelling,

Snowball Sampling and Citation Analysis. *CEUR-WS*, 2018. Vol. 2105. Pp. 179–192.

9. Lecy J., Beatty K. Representative Literature Reviews Using Constrained Snowball Sampling and Citation Network Analysis. 2012. *SSRN Electronic Journal*, 2012.

10. Ahad A., Fayaz M., Shah A. Navigation through Citation Network Based on Content Similarity Using Cosine Similarity Algorithm. *International Journal of Database Theory and Application*. 2016. Vol. 9, No. 5. Pp. 9–20.

11. Schьtze H., Manning C. D., Raghavan P. *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press, 2008.

12. Batagelj V. Efficient algorithms for citation network analysis. 2003. URL: https://arxiv.org/abs/cs/0309023.

13. Tatarintseva O., Ermolayev V., Keller B., Matzke W.-E. Quantifying ontology fitness in OntoElect using saturation- and vote-based metrics. / eds. Ermolayev V., Mayr H.C., Nikitchenko M., Spivakovsky A., Zholtkevych G. Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2013. Springer, Cham, 2013. *Communications in Computer and Information Science*. Vol. 412. Pp. 136–162.

14. Dobrovolskyi H., Keberle N. Probabilistic topic modelling for controlled snowball sampling in citation network collection. Proceedings of the International Conference on Knowledge Engineering and the Semantic Web. Springer, 2017. Pp. 85–100.

15. Yan X., Guo J., Lan Y., Cheng X. A biterm topic model for short texts. Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013. Pp. 1445–1456.

16. Zuo Y., Zhao J., Xu K. Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 2016. Vol. 48, No. 2. Pp. 379–398.

17. Choi S.S., Cha S.H., Tappert C.H. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*. 2010. Vol. 8, No. 1, Pp. 43–48.

18. Lance G., Williams W. A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. *The Computer Journal*. 1967. Vol. 9, No. 4. Pp. 373–380.

19. Frantzi K., Ananiadou S. The C-value/NC-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing*. 1999. Vol. 6, No. 3. Pp. 145–179.

MATHEMATICAL MODEL TO SELECT THE SCIENTIFIC PUBLICATION IN THE PROCESS OF BIBLIOGRAPHICAL INDEX COMPILATION

H. Dobrovolskyi, N. Keberle

Zaporizhzhya National University

vul. Zhukovskogo, 66, Zaporizhzhya, 69600, Ukraine. E-mail: gen.dobr@gmail.com

The purpose of the paper is to present a hybrid mathematical model of the process of bibliographic discovery and bibliographic selection of the scientific publications. The model describes the process of meeting an information need for the recommendation or scientific-supporting bibliographical index. The index must consist only of documents belonging to a selected topic where the documents are seminal ones and contain all basic statements concerning the topic. At the same time the index must be as short as possible allowing detailed analysis of all the documents and appending them to short scientific paper as references. **Methodology**. The presented study assumes that an index creator has too low expertise in the topic to create ideal set of keywords however his/her experience allows correct relevance estimation if document is found. The guessed necessary condition of the presence of all the basic statements concerning the topic is the terminological saturation of the selected ordered document set. The model is built upon consistent formal description of the controlled snowball method developed by the authors. The model's formal description uses language of the mathematical statistics, graph theory and set theory. The main components of the presented model are probabilistic topic model, mapping generating the controlled snowball iterations, conditions of a fixed iteration point, paths analysis in the documents citation network, and automatic terms detection. **Originality.** The developed model differs from existing ones by consistent combination of the mentioned components, application of the sparse symmetric nonnegative matrix factorization to build the topic model, existence of the fixed iteration point conditions and the condition of the terminological saturation of the selected ordered document set. The model provides solid base for the method of bibliographic discovery and selection, developed in earlier works, and corresponding information technology. **Findings.** Main results of the study are: definition of the recommendation and scientific-supporting bibliographic index as a minimal terminologically saturated ordered document set; definition of the quality measure as the size of the set. Simplified iteration convergence criteria are proposed as well. **Practical value** of the results is the application of the developed model to automatic collection and renewal of the personalized publication index that contains all seminal results of the selected topic of scientific knowledge and at the same time is short enough to be analyzed in details. Indexes collected in such a way can be used to survey the state-of-the-art in topic of interest that is important part of scientific research. **Conclusions**. Formal mathematical description of the method, developed in earlier works, allows clear circumvention of its applicability domain, introduces quantitative quality measures and allows to plan the future investigations to elaborate model stability and connection of the simplifies convergence criteria and terminological saturation.

Key words: bibliography, selection, citation network, topic model, terminology detection, quality measure, terminological saturation.

REFERENCES

1. Halhanova, O. (2015), Skladannya bibliografichnykh posibnykiv: inform.-metod. Materialy [Compilaiton of bibliographical indexes: information and methodical materials], Kyiv, Ministry of Culture of Ukraine.

2. Fisch, C., Block, J. (2018), "Six tips for your (systematic) literature review in business and management research", *Management Review Quarterly*, 68(2), pp. 103-106.

3. Petticrew, M., Gilbody, S. (2004), "Planning and conducting systematic reviews", *Health psychology in practice*. pp. 150-179.

4. Friday, D., Ryan, S., Sridharan, R., Collins, D. (2018), ollaborative risk management: a systematic literature review", *International Journal of Physical Distribution & Logistics Management*, 48(3), pp. 231-253.

5. Colicchia, C., Strozzi, F. (2012), "Supply chain risk management: a new methodology for a systematic literature review", *Supply Chain Management: An International Journal*, 17(4), pp.403-418.

6 Varela, A., Pratt, M., Harris, J., Lecy, J., Salvo, D., Brownson, R., Hallal, P. (2018), "Mapping the historical development of physical activity and health research: A structured literature review and citation network analysis", *Preventive Medicine*, 111, pp. 466-472.

7 Liu, J., Lu, L., Lu, W., Lin, B. (2013), "Data envelopment analysis 1978–2010: A citation-based literature survey", *Omega*, 41(1), pp. 3-15.

8. Dobrovolskyi, H., Keberle, N. (2018), Collecting the Seminal Scientific Abstracts with Topic Modelling, Snowball Sampling and Citation Analysis. CEUR-WS, vol. 2105, pp. 179-192.

9. Lecy, J., Beatty, K. (2012), "Representative Literature Reviews Using Constrained Snowball Sampling and Citation Network Analysis", *SSRN Electronic Journal*, 2012.

10. Ahad, A., Fayaz, M., Shah, A. (2016), "Navigation through Citation Network Based on Content Similarity Using Cosine Similarity Algorithm", *International Journal of Database Theory and Application*, 9(5), pp. 9-20.

11. Schitze, H., Manning, C. D., Raghavan, P. (2008), Introduction to information retrieval, National Cambridge University Press, Cambridge, UK.

12. Batagelj, V. (2003), "Efficient algorithms for citation network analysis", *Archiv preprint cs/0309023*. URL: https://arxiv.org/abs/cs/0309023.

13. Tatarintseva, O., Ermolayev, V., Keller, B., Matzke, W. E. (2013), "Quantifying ontology fitness in OntoElect using saturation-and vote-based metrics", *International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications*, Springer, Vol. 412, pp. 136-162.

14. Dobrovolskyi, H., Keberle, N., Todoriko, O. (2017), "Probabilistic topic modelling for controlled snowball sampling in citation network collection", *International Conference on Knowledge Engineering and the Semantic Web*, Proceedings of the International Conference on Knowledge Engineering and the Semantic Web, Springer, pp. 85-100.

15. Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013),. "A biterm topic model for short texts", *Proceedings of the 22nd international conference on World Wide Web*, ACM, pp. 1445-1456.

16. Zuo, Y., Zhao, J., Xu, K. (2016), "Word network topic model: a simple but general solution for short and imbalanced texts", *Knowledge and Information Systems*, 48(2), pp. 379- 398.

17. Choi, S. S., Cha, S. H., Tappert, C. C. (2010), "A survey of binary similarity and distance measures". *Journal of Systemics, Cybernetics and Informatics*, 8(1), pp. 43-48.

18. Lance, G. N., Williams, W. T. (1967), "A general theory of classificatory sorting strategies: 1. Hierarchical systems", *The Computer Journal*, 9(4), pp. 373-380.

19. Frantzi, K. T., Ananiadou, S. (1999), "The C-value/NC-value domain-independent method for multi-word term extraction", *Journal of Natural Language Processing*, 6(3), pp. 145-179.

Стаття надійшла 23.12.2019.