УДК 519.254+001.2
DOI: 10.20535/SRIT.2308-8893.2021.2.09

ТЕМАТИЧНЕ МОДЕЛЮВАННЯ НАУКОВЦІВ НА ОСНОВІ ЇХ ІНТЕРЕСІВ У GOOGLE SCHOLAR

С.Д. ШТОВБА, М.В. ПЕТРИЧКО

Анотація. Запропоновано алгоритм тематичного моделювання науковців за науковими спеціальностям на основі їх інтересів у профілях у Google Scholar. Алгоритм використовує перелік наукових спеціальностей із системи класифікації наук ANZSRC. Інформаційним ресурсом для тематичного моделювання є база категоризованих наукових публікацій із системи Dimensions. Інтереси з профілів науковців використовуються як пошукові запити для Dimensions, сервіси якої видають розподіли релевантних документів за спеціальностями. Для зменшення інформаційного шуму ці розподіли проходять декілька етапів оброблення. Порівнюються результати тематичного моделювання на основі профільних інтересів у Google Scholar і категоризованого списку авторських публікацій у Dimensions за метрикою Чекановського з урахуванням спорідненості спеціальностей. Для тестових науковців виявлено високу узгодженість результатів тематичного моделювання за різної початкової інформації.

Ключові слова: тематичне моделювання, категоризація, Google Scholar, Dimensions, ANZSRC, профіль науковця, наукові інтереси, метрика Чекановського, індекс Жакара.

ВСТУП

Сьогодні професійні спільноти людей взаємодіють у різноманітних онлайнових мережах. Не винятком є і спільнота науковців. Найбільшою онлайновою мережею науковців є Google Scholar. Зокрема, у цій мережі у відкритому доступі понад 50 тисяч профілів українських науковців. Такий величезний ресурс виглядає привабливим для розроблення технологій аналітичного опрацювання нагромадженої в ньому інформації з метою ідентифікації лідерів — статей, науковців, університетів та журналів; виявлення тенденцій наукових досліджень; кластеризації науковців; підбору партнерів для спільних проєктів, опонентів дисертацій, рецензентів рукописів тощо.

Найчастіше із профілів науковців у Google Scholar використовують дані про цитованість. Її, наприклад, використовують як початкові дані для рейтингування університетів у Webometrics. Створено також кілька інформаційних систем на базі Google Scholar, найбільш відомими серед яких є Publish or Perish і Scholarometer [1]. Багато досліджень, зокрема [2, 3], стосуються перевірки достовірності цитованості в Google Scholar порівняно з наукометричними системами Scopus, Web of Science, Dimensions та іншими, які наповнюються виключно за метаданими з видавництв.

Окрім списку публікацій та їх цитування у профілі науковця міститься і інша інформація. Зокрема, науковець у профілі вказує свої інтереси, і робить він це на власний розсуд, обираючи слова у довільний спосіб. Google Scholar дозволяє виконувати пошук науковців за тим чи іншим інтересом. Але видачі формуються за буквальним збігом. Тому видачі для *fuzzy set* і *fuzzy sets*

будуть різними, не говорячи вже про синонімічні інтереси типу *fuzzy evidence* і *fuzzy inference*. Google Scholar не враховує і сукупність інтересів користувача, тобто пошук за кожним інтересом виконується незалежно та ізольовано. Відповідно, в одну видачу потраплять науковці, що поміж своїх інтересів вказали *безпеку* в сенсі як *security*, так і *safety*. Таким чином, пошукові та аналітичні сервіси за велетенським масивом профілів науковців у Google Scholar досить примітивні.

Мета роботи — тематичне моделювання науковців на основі їх інтересів у Google Scholar. Методи опрацювання інтересів із профілів науковців у Google Scholar є мало дослідженими. Нами виявлено лише дві релевантні публікації. Перша з них [4] стосується рекомендаційної системи для підбору наукового керівника, яка поміж інших джерел інформації використовує і інтереси кандидатів з їх профілів у Google Scholar. Друга публікація [5] описує інформаційну технологію синтезу наукового профілю інституту чи дослідницької лабораторії. Ця технологія, поміж іншої інформації, використовує і інтереси науковців з їх профілів у Google Scholar. Праці [4, 5] базуються на використанні попарного порівняння за косинусової метрики близькості — відстані між науковцем та набором ключових слів з деякої тематики. Такою тематикою в [4] обрано статтю у Вікіпедії. На відміну від цих праць, будемо намагатися категоризувати науковців у межах деякої класифікації наук, тобто розподілити їх за науковими спеціальностями.

Автоматична категоризація науковців виконується зазвичай в результаті узагальнення тематик їх публікацій. Для цього у праці [6] запропоновано статистичну модель «автор – тема» на основі тематичного моделювання з використанням прихованого розподілу Діріхле (LDA) [7]. Модель подає науковця як розподіл над деякими абстрактними темами. Теми є кластерами схожих слів. Її недоліком є погана інтерпретація тем, оскільки вони формуються за частотою слів у одному документі. Для покращення інтерпретації у [8] запропоновано модель «автор – дисципліна – тема». У ній для опису науковця додатково використовують наукову спеціальність, яка визначається за журналом чи збірником статей, у якому опубліковано аналізовану працю. На виході науковець подається сукупністю належностей до наукових спеціальностей. У праці [9] для підбору рецензентів запропоновано модель «автор – персона – тема». У ній враховано те, що автори часто пишуть про декілька різних комбінацій тем з однієї предметної галузі. Дуже рідко особа є експертом в усіх аспектах якоїсь предметної галузі. За результатами моделювання науковця зіставляють з декількома персонами (personas). Кожна персона є кластером статей науковця зі своїм тематичним розподілом. У праці [10] розвинуто методи [8, 9] моделлю «автор – інтерес – тема», яка містить документи зі схожими темами як один клас документів, подібно до того, як тематичні моделі подають спільну появу (co-occurrence) слів як одну тематичну змінну.

Окрім методів на основі тематичного моделювання також використовуються моделі на основі ембедингу слів (word embedding) [11–14]. Однією з найпопулярніших моделей ембедингу слів є модель word2vec [15]. На відміну від прихованого розподілу Діріхле [7], прихованого семантичного аналізу (pLSA) [16] та інших статистичних моделей, які породжують імовірнісний розподіл на основі спільної появи слів та документів, word2vec

фокусується на контекстуальній (семантичній та синтаксичній) інформації слів. Згадані методи показують непогану ефективність для таких завдань, як рекомендація рецензентів, пошук експертів тощо. Результати моделювання подаються у вигляді векторів, які складно інтерпретувати.

Проаналізовані методи передбачають наявність достатньої кількості статей науковця з виділеними ключовими словами. При цьому не враховується, що співавторами статті можуть бути кілька науковців, на кожного з яких припадає деяка підмножина з усього списку ключових слів. Причому з десятка ключових слів статті внесок співавтора може відображати лише одне ключове слово. Крім того, науковець, особливо молодий, може і не мати достатньої кількості статей для достовірної категоризації. Утім він може самостійно задати у профілі набір ключових слів, який описує його дослідження. Із часом науковець може змінити напрям своєї діяльності, наприклад, працювати в іншій лабораторії чи над іншим проєктом. Але його продовжуватимуть категоризувати за давніми публікаціями. У зв'язку з цим виникла зацікавленість у тематичному моделюванні на основі інтересів, які науковець власноруч сформулював на поточний момент, тобто на основі актуальної та узагальненої початкової інформації, що позбавлена наведених вище недоліків.

ПОСТАНОВКА ЗАДАЧІ

Вважатимемо відомими:

$W = (w_1, w_2, ..., w_n)$ — список ключових слів, якими науковець у своєму профілі в Google Scholar на власний розсуд описав свої інтереси;

$T = (t_1, t_2, ..., t_m)$ — перелік можливих тем у формі списку наукових спеціальностей за деякою класифікацією наук;

$D_1, D_2, ..., D_m$ — тематичні колекції розмічених текстів, кожна з яких містить лише публікації з тем $t_1, t_2, ..., t_m$ відповідно;

$B = D_1 \cup D_2 \cup ... \cup D_m$ — загальна колекція розмічених текстів, тобто множина публікацій, кожна з яких стосується однієї або декількох тем з множини T;

$R(D,T) \subset D \times T$ — відношення, яке описує належність публікацій до тематичних колекцій.

Задача полягає у знаходженні тем з T, яким відповідає сукупність інтересів W. Будемо вказувати не лише сам факт належності, але і ступінь належності. Таким чином, на виході отримуємо нечітку множину $\widetilde{W}$ на універсальній множині тем T:

$$\widetilde{W} = \left(\frac{\mu_W(t_1)}{t_1}, \frac{\mu_W(t_2)}{t_2}, ..., \frac{\mu_W(t_m)}{t_m} \right),$$

де $\mu_W(t_p) \in [0,1]$ — ступінь належності сукупності інтересів W до спеціальності t_p, $p = \overline{1,m}$.

На $\widetilde{W}$ накладемо такі обмеження:

1) потужність носія нечіткої множини $\widetilde{W}$ має бути невеликою $1 \le \left|\text{support}(\widetilde{W})\right| \le T_{\max}$, наприклад, за $T_{\max} \in \{2, 3, 4\}$ науковець відповідатиме лише кільком спеціальностям;

2) $\sum_{p=\overline{1,m}} \mu_W(t_p) = 1$, що ототожнюється з умовою регуляризації тематичного моделювання.

ДОБУВАННЯ ПОЧАТКОВИХ ДАНИХ

Для отримання списку ключових слів науковця скористаємося його профілем у Google Scholar. Для прикладу на рис. 1 наведено профіль науковця з двома ключовими словами $w_1 =$"neural networks" та $w_2 =$"artificial intelligence". Послідовність ключових слів у множині W неважлива, що відповідає врахуванню інформації за схемою мішка слів (bag of words). Часто інтереси у профілі доповнюють один одного, тим самим фокусуючи тематику досліджень. Щоб це врахувати синтезуємо додаткові ключові слова у вигляді пар початкових інтересів. Інтереси в парах поєднаємо логічною операцією *TA*. Для науковця з рис. 1 додаткове ключове слово запишемо як $w_3 =$"neural networks" AND "artificial intelligence". Якщо у профілі науковця вказано три інтереси, буде синтезовано три додаткові ключові слова, якщо у профілі чотири інтереси, тоді синтезується шість додаткових ключових слів, якщо п’ять інтересів, тоді десять додаткових ключових слів тощо. Синтез додаткових парних ключових слів є своєрідним аналогом дистантного поєднання слів (word co-occurrence), яке дозволяє зменшити вербальний шум.

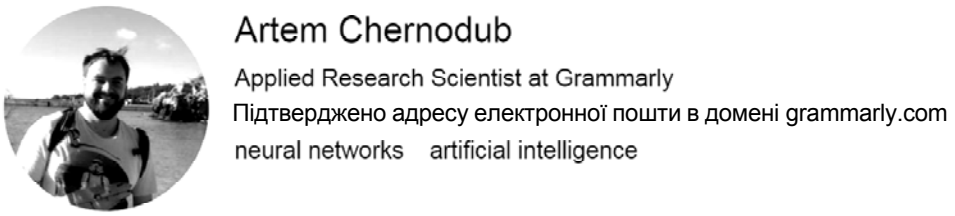

Рис. 1. Приклад профілю науковця з двома інтересами

Для тематичного моделювання науковців необхідно обрати систему класифікації наукових спеціальностей. Їх багато, але під час вибору системи класифікації врахуємо не лише її змістовні переваги і недоліки, але і наявність відповідної інформаційної системи з доступними пошуковими сервісами. При цьому база даних системи має індексувати велику кількість категоризованих публікацій, які охоплюють усі наукові галузі. Інформаційною системою, яка задовольняє перераховані вимоги, є Dimensions.

Натепер Dimensions індексує понад 110 млн публікацій. Усі публікації в Dimensions категоризовано за дворівневим варіантом Австралійсько-новозеландського стандарту ANZSRC (Australian and New Zealand Standard Research Classification). У ньому науку поділено на 22 галузі (Divisions) із 154 спеціальностями (Research Groups). Цей дворівневий варіант ANZSRC, який і будемо надалі використовувати, подано в табл. 1.

Таблиця 1. Система класифікації наук ANZSRC, що використовується у Dimensions

Галузь	Спеціальність
Mathematical Sciences	A1 – Pure Mathematics; A2 – Applied Mathematics; A3 – Numerical and Computational Mathematics; A4 – Statistics; A5 – Mathematical Physics
Physical Sciences	B1 – Astronomical and Space Sciences; B2 – Atomic, Molecular, Nuclear, Particle and Plasma Physics; B3 – Classical Physics; B4 – Condensed Matter Physics; B5 – Optical Physics; B6 – Quantum Physics; B7 – Other Physical Sciences
Chemical Sciences	C1 – Analytical Chemistry; C2 – Inorganic Chemistry; C3 – Macromolecular and Materials Chemistry; C4 – Medicinal and Biomolecular Chemistry; C5 – Organic Chemistry; C6 – Physical Chemistry (incl. Structural); C7 – Theoretical and Computational Chemistry; C8 – Other Chemical Sciences
Earth Sciences	D1 – Atmospheric Sciences; D2 – Geochemistry; D3 – Geology; D4 – Geophysics; D5 – Oceanography; D6 – Physical Geography and Environmental Geoscience; D7 – Other Earth Sciences
Environmental Sciences	E1 – Ecological Applications; E2 – Environmental Science and Management; E3 – Soil Sciences; E4 – Other Environmental Sciences
Biological Sciences	F1 – Biochemistry and Cell Biology; F2 – Ecology; F3 – Evolutionary Biology; F4 – Genetics; F5 – Microbiology; F6 – Physiology; F7 – Plant Biology; F8 – Zoology; F9 – Other Biological Sciences
Agricultural and Veterinary Sciences	G1 – Agriculture, Land and Farm Management; G2 – Animal Production; G3 – Crop and Pasture Production; G4 – Fisheries Sciences; G5 – Forestry Sciences; G6 – Horticultural Production; G7 – Veterinary Sciences; G8 – Other Agricultural and Veterinary Sciences
Information and Computing Sciences	H1 – Artificial Intelligence and Image Processing; H2 – Computation Theory and Mathematics; H3 – Computer Software; H4 – Data Format; H5 – Distributed Computing; H6 – Information Systems; H7 – Library and Information Studies; H8 – Other Information and Computing Sciences
Engineering	I1 – Aerospace Engineering; I2 – Automotive Engineering; I3 – Biomedical Engineering; I4 – Chemical Engineering; I5 – Civil Engineering; I6 – Electrical and Electronic Engineering; I7 – Environmental Engineering; I8 – Food Sciences; I9 – Geomatic Engineering; I10 – Manufacturing Engineering; I11 – Maritime Engineering; I12 – Materials Engineering; I13 – Mechanical Engineering; I14 – Resources Engineering and Extractive Metallurgy; I15 – Interdisciplinary Engineering; I16 – Other Engineering
Technology	J1 – Agricultural Biotechnology; J2 – Environmental Biotechnology; J3 – Industrial Biotechnology; J4 – Medical Biotechnology; J5 – Communications Technologies; J6 – Computer Hardware; J7 – Nanotechnology; J8 – Other Technology
Medical and Health Sciences	K1 – Medical Biochemistry and Metabolomics; K2 – Cardiorespiratory Medicine and Haematology; K3 – Clinical Sciences; K4 – Complementary and Alternative Medicine; K5 – Dentistry; K6 – Human Movement and Sports Science; K7 – Immunology; K8 – Medical Microbiology; K9 – Neurosciences; K10 – Nursing; K11 – Nutrition and Dietetics; K12 – Oncology and Carcinogenesis; K13 – Ophthalmology and Optometry; K14 – Paediatrics and Reproductive Medicine; K15 – Pharmacology and Pharmaceutical Sciences; K16 – Medical Physiology; K17 – Public Health and Health Services; K18 – Other Medical and Health Sciences
Built Environment and Design	L1 – Architecture; L2 – Building; L3 – Design Practice and Management; L4 – Engineering Design; L5 – Urban and Regional Planning; L6 – Other Built Environment and Design
Education	M1 – Education Systems; M2 – Curriculum and Pedagogy; M3 – Specialist Studies In Education; M4 – Other Education
Economics	N1 – Economic Theory; N2 – Applied Economics; N3 – Econometrics; N4 – Other Economics
Commerce, Management, Tourism and Services	O1 – Accounting, Auditing and Accountability; O2 – Banking, Finance and Investment; O3 – Business and Management; O4 – Commercial Services; O5 – Marketing; O6 – Tourism; O7 – Transportation and Freight Services

Продовження табл.. 1

Галузь	**Спеціальність**
Studies in Human Society	P1 – Anthropology; P2 – Criminology; P3 – Demography; P4 – Human Geography; P5 – Policy and Administration; P6 – Political Science; P7 – Social Work; P8 – Sociology; P9 – Other Studies In Human Society
Psychology and Cognitive Sciences	Q1 – Psychology; Q2 – Cognitive Sciences; Q3 – Other Psychology and Cognitive Sciences
Law and Legal Studies	R1 – Law; R2 – Other Law and Legal Studies
Studies in Creative Arts and Writing	S1 – Art Theory and Criticism; S2 – Film, Television and Digital Media; S3 – Journalism and Professional Writing; S4 – Performing Arts and Creative Writing; S5 – Visual Arts and Crafts; S6 – Other Studies In Creative Arts and Writing
Language, Communication and Culture	T1 – Communication and Media Studies; T2 – Cultural Studies; T3 – Language Studies; T4 – Linguistics; T5 – Literary Studies; T6 – Other Language, Communication and Culture
History and Archaeology	U1 – Archaeology; U2 – Curatorial and Related Studies; U3 – Historical Studies; U4 – Other History and Archaeology
Philosophy and Religious Studies	V1 – Applied Ethics; V2 – History and Philosophy of Specific Fields; V3 – Philosophy; V4 – Religion and Religious Studies; V5 – Other Philosophy and Religious Studies

Запит до інформаційної системи Dimensions формуємо окремо за кожним елементом множини *W*. Якщо цей елемент є словосполученням, тоді подамо його у лапках. Пошук виконуємо за назвою та рефератом публікацій 2016–2020 рр. Приклад видачі за пошуковим запитом *"neural networks"* подано у вигляді рис. 2. За кожною спеціальністю та за кожною галуззю виводиться кількість публікацій, у назві або в рефераті яких фігурує пошуковий вираз. Видачу відсортовано за спаданням кількості публікацій. Також можна отримати загальну кількість публікацій за кожною спеціальністю, тобто обсяги тематичних колекцій.

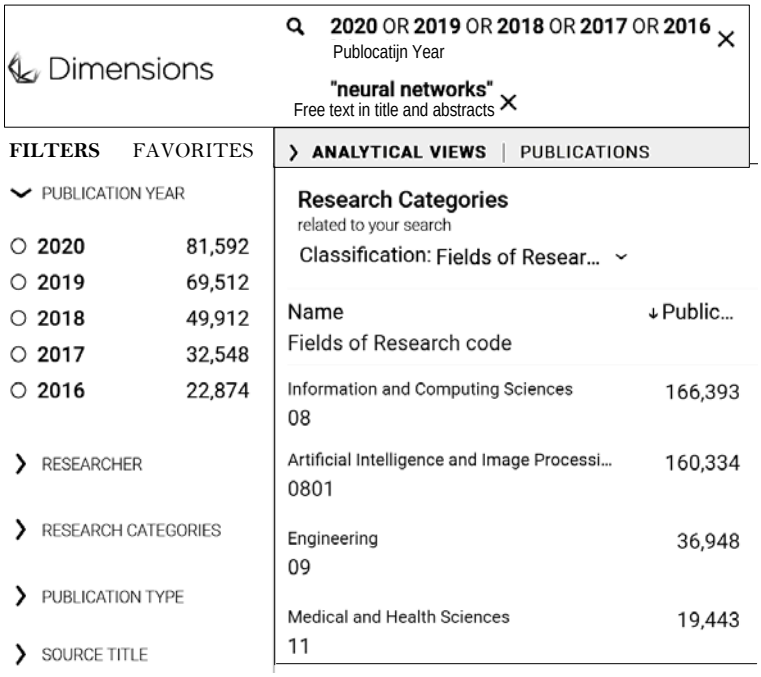

Рис. 2. Видача Dimensions за пошуковим запитом "*neural networks*"

Dimensions індексує переважно англомовні публікації, тому всі інтереси з профілю науковця в Google Scholar необхідно попередньо перекласти англійською мовою. Інколи науковці вказують у своєму профілі один і той самий інтерес кількома мовами, наприклад, *neural networks* та *нейронні мережі*. У такому випадку ці два інтереси об'єднаємо в один англомовний — *neural networks*.

АЛГОРИТМ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ

Тематичне моделювання науковців виконаємо на базі таких принципів:

- *статистичного підтримання* — чим більша частка публікацій з певної спеціальності містить аналізоване ключове слово, тим більша належність ключового слова до цієї спеціальності;
- *багатоярликовості* — ключове слово може належати до кількох спеціальностей;
- *фільтрації шумів* — ігноруються спеціальності, до яких ключове слово належить з незначним ступенем;
- *ігнорування стоп-слів* — ігнорується ключове слово, яке трапляється у дуже багатьох категоризованих публікаціях;
- *солідарності* — чим більше ключових слів за окремими запитами належить до однієї і тієї ж спеціальності, тим більша можливість належності науковця до цієї спеціальності;
- *фокусування* — якщо в тематичній колекції багато публікацій, які містять кілька ключових слів науковця одночасно, тоді збільшуються шанси належності науковця до відповідної спеціальності.
- *компактності* — один науковець може належати лише до невеликої кількості спеціальностей;
- *взаємодії спеціальностей* — під час відсікання хвоста розподілу тем, внесок мінорних спеціальностей перерозподіляється на лідерів з урахуванням їх схожості.

Наведені принципи пропонується реалізувати алгоритмом, який містить три ділянки. На першій ділянці формується множина запитів на основі ключових слів та їх поєднання. Використовуємо лише пари ключових слів, тому що видачі за трійками часто виявляються порожніми, але при цьому суттєво збільшується тривалість пошуку.

На другій ділянці алгоритму (рис. 3) виконується тематичне моделювання за кожним запитом окремо. Спеціальності обираємо за частотою входження запиту в тематичну колекцію. Частота розраховується як відношення кількості документів, що містять пошуковий вираз, до загальної кількості документів зі спеціальності. При цьому стоп-слова та шуми фільтруються за кількістю входжень в усю колекцію документів із застосуванням порогових значень. Вилучаються і мінорні спеціальності. Спочатку вилучаємо за пороговим значенням кількості знайдених документів, які належать до відповідної спеціальності, а потім — за кумулятивним принципом, відсікаючи хвіст розподілу за пороговим значенням.

На третій ділянці алгоритму (рис. 4) усереднюємо належності за всіма запитами та відсікаємо хвіст сукупного розподілу за пороговим значенням. Далі вилучаємо спеціальності з низьким рівнем належності таким чином, щоб результат став компактним і представницьким, коли науковець відповідає не більше ніж чотирьом спеціальностям, причому до кожної з них належність є значущою.

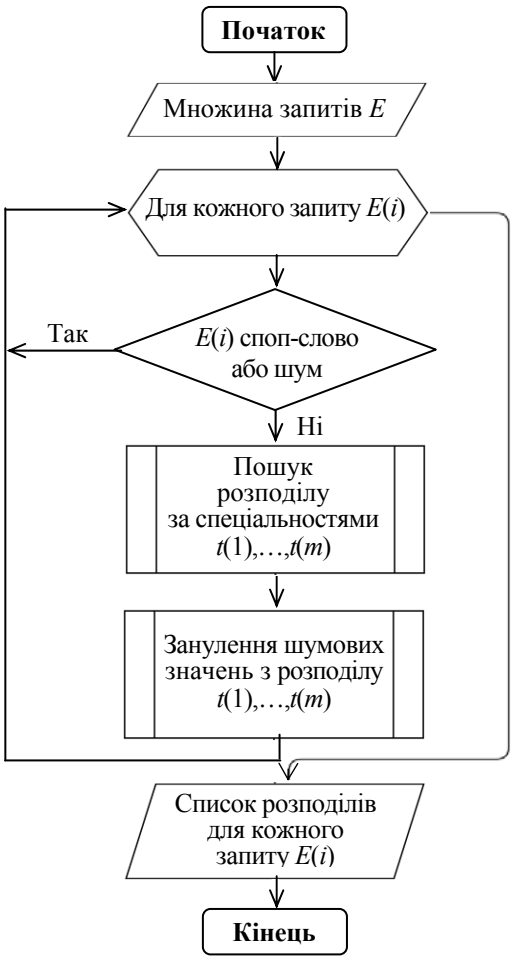

Рис. 3. Блок-схема другої ділянки алгоритму тематичного моделювання

Рис. 4. Блок-схема третьої ділянки алгоритму тематичного моделювання

На третій ділянці роботи алгоритму під час ітераційного вилучення мінорної спеціальності її внесок перерозподіляється на інші спеціальності з урахуванням коефіцієнтів схожості із праці [17]. Наприклад, нехай на проміжному етапі науковця віднесено до наукових спеціальностей таким чином: $\widetilde{W} = \left(\frac{0{,}5}{\text{Н6}}, \frac{0{,}2}{\text{О5}}, \frac{0{,}2}{\text{О6}}, \frac{0{,}1}{\text{О4}} \right)$. Вилучимо мінорну спеціальність О4. Для цього спочатку за методом [17] знайдемо коефіцієнти Жакара між О4 та іншими спеціальностями. Вони за даними 2016–2020 рр. такі: $J(\text{О4}, \text{Н6}) = 0$, $J(\text{О4}, \text{О5}) = 0{,}13$, $J(\text{О4}, \text{О6}) = 0{,}22$. З урахуванням схожості внесок мінорної спеціальності О4 перерозподіляється таким чином:

$\widetilde{W} = \left(\frac{0{,}5 + 0 \cdot 0{,}1}{\text{Н6}}, \frac{0{,}2 + 0{,}13 \cdot 0{,}1}{\text{О5}}, \frac{0{,}2 + 0{,}22 \cdot 0{,}1}{\text{О6}} \right)$. Підрахувавши, отримуємо $\widetilde{W} = \left(\frac{0{,}5}{\text{Н6}}, \frac{0{,}213}{\text{О5}}, \frac{0{,}222}{\text{О6}} \right)$. Після нормування на 1 маємо результат: $\widetilde{W} = \left(\frac{0{,}535}{\text{Н6}}, \frac{0{,}228}{\text{О5}}, \frac{0{,}237}{\text{О6}} \right)$.

ПОКРОКОВИЙ КОНТРОЛЬНИЙ ПРИКЛАД

Проілюструємо роботу алгоритму на прикладі тематичного моделювання науковця з рис. 1. За двома інтересами науковця сформовано три пошукові запити. Частоту входжень трьох ключових слів у тематичні колекції показано на рис. 5, а результати після першого відсікання хвостів розподілів — на рис. 6. Далі усереднюємо за усіма запитами (рис. 7) і відсікаємо хвіст розподілу (рис. 8). Проміжний розподіл став перенаповненим через заширо-ке формулювання науковцем своїх інтересів. Для фокусування результатів тематичного моделювання на заключному етапі алгоритму розподіл обрізаємо до двох спеціальностей (рис. 9). У результаті науковець з інтере-сами в галузі штучного інтелекту та нейронних мереж найбільше відповідає спеціальностям *H1 – Artificial Intelligence and Image Processing* зі ступенем належності 0,767 та *Q2 – Cognitive Sciences* зі ступенем належності 0,233. Така категоризація науковця не суперечить поглядам авторів цієї статті. Із прикладу видно, що навіть за двома початковими ключовими словами за-пропонований алгоритм достатньо точно знаходить відповідність науковця спеціальностям.

ПОРІВНЯННЯ З КАТЕГОРИЗАЦІЄЮ ЗА СТАТТЯМИ

Перевіримо узгодженість результатів тематичного моделювання науковців на основі ключових слів з їх профілів у Google Scholar та на основі катего-ризованих статей у Dimensions. Для цього відберемо трьох науковців: А. Чернодуба (див. рис 1), Є. Бодянського (рис. 10) та Н. Куссуль (рис. 11). Ці науковці мають у Dimensions велику кількість публікацій за п’ять остан-ніх років, що дозволяє отримати статистично значущі результати.

За аналізований період А. Чернодуб опублікував 22 праці, які категори-зовано за п’ятьма спеціальностями. Найбільше публікацій — 11 потрапило до спеціальності *H1*. Є. Бодянський опублікував 88 робіт. Вони категоризо-вані за 12 спеціальностями. Найбільше публікацій — 59 потрапило до спеціальності *H1*. Н. Куссуль опублікувала 47 робіт, які категоризовано за 14 спеціальностями. Найбільше публікацій — 21 потрапило до спеціаль-ності *I9*.

За розподілами публікацій за спеціальностями з використанням третьої ділянки алгоритму тематичного моделювання отримаємо належності нау-ковців до спеціальностей (табл. 2). Там же вказано результати тематичного моделювання на основі інтересів науковців у Google Scholar.

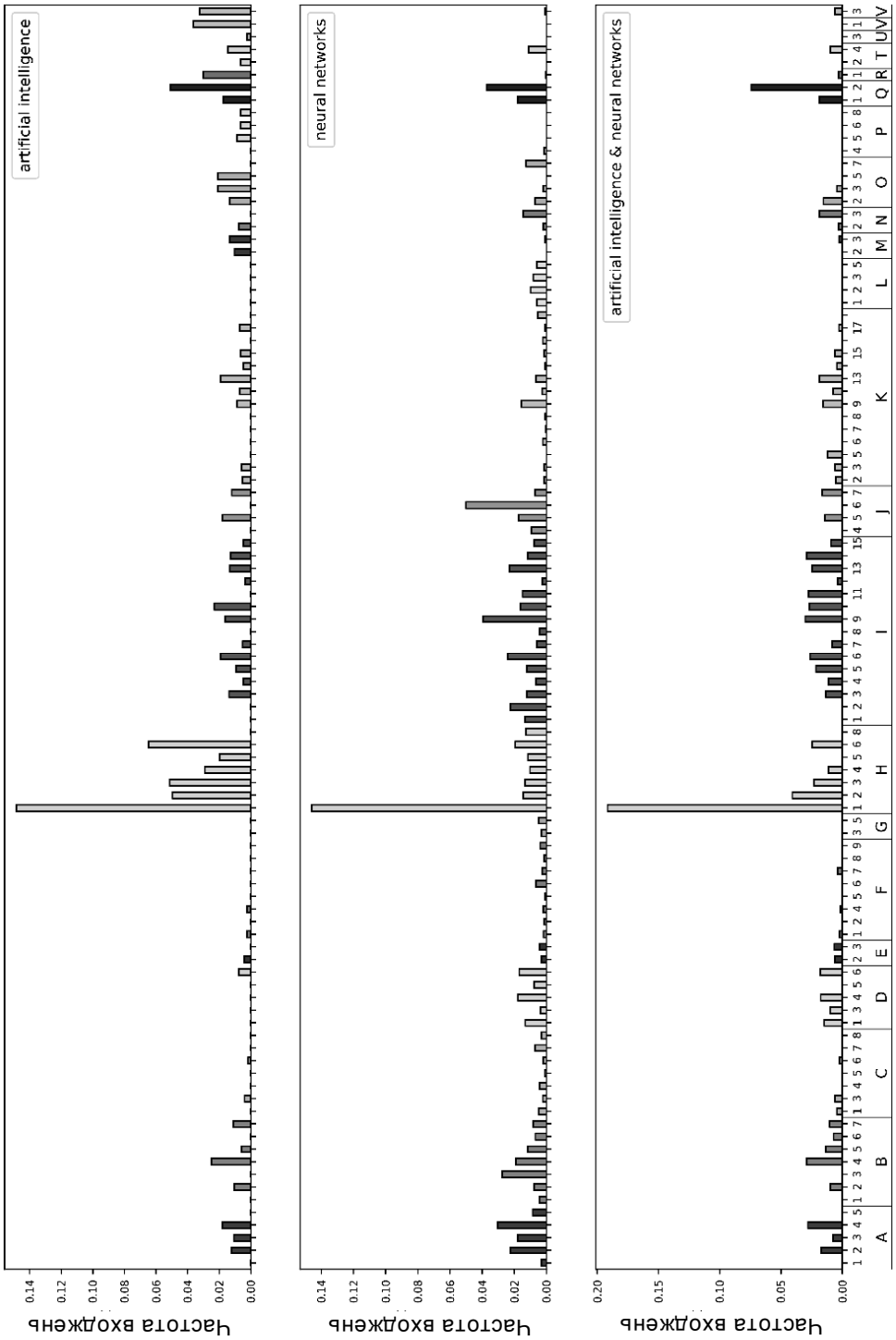

Рис. 5. Початковий розподіл належності кожного інтересу до спеціальностей

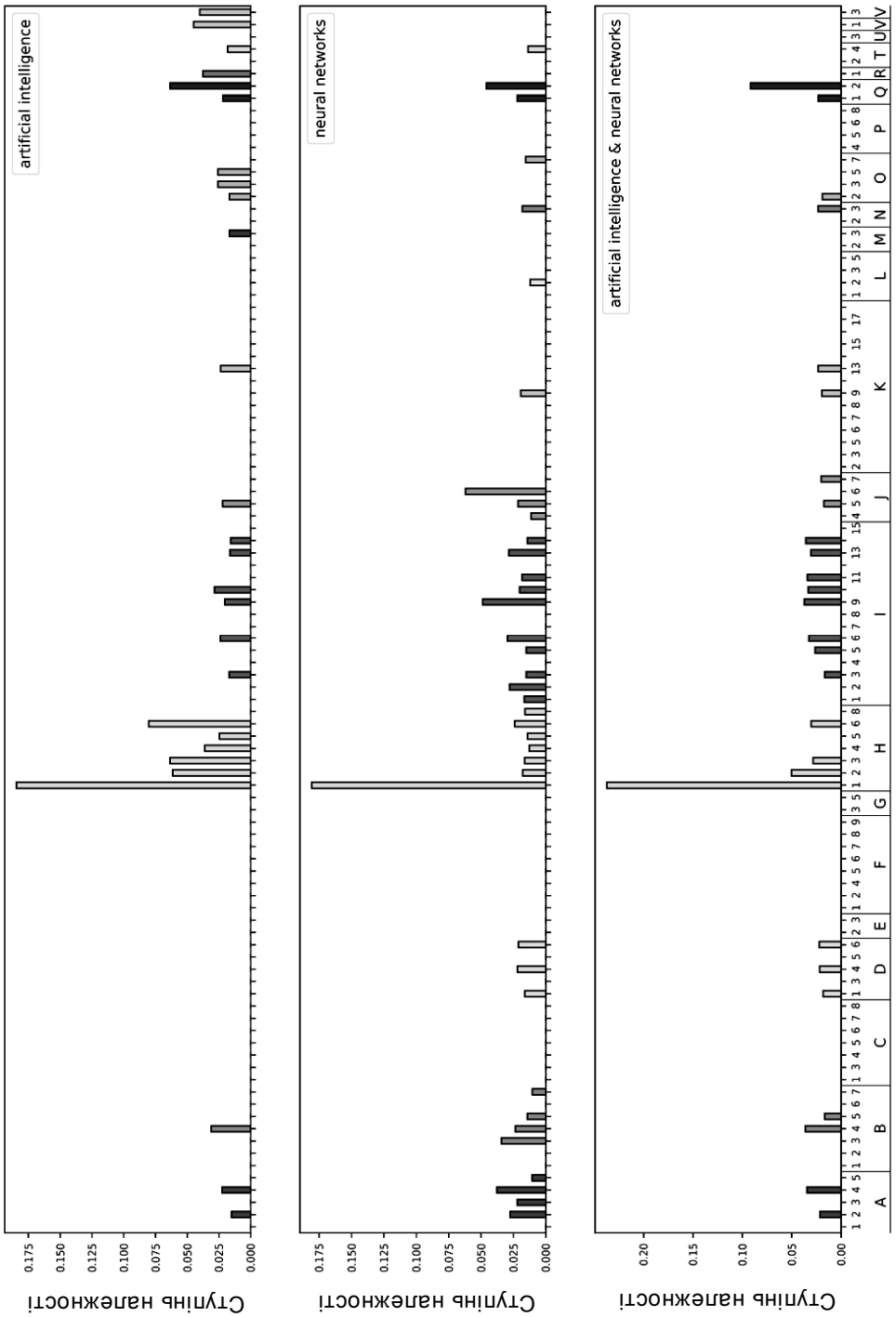

Рис. 6. Проріджені розподіли після першої фільтрації

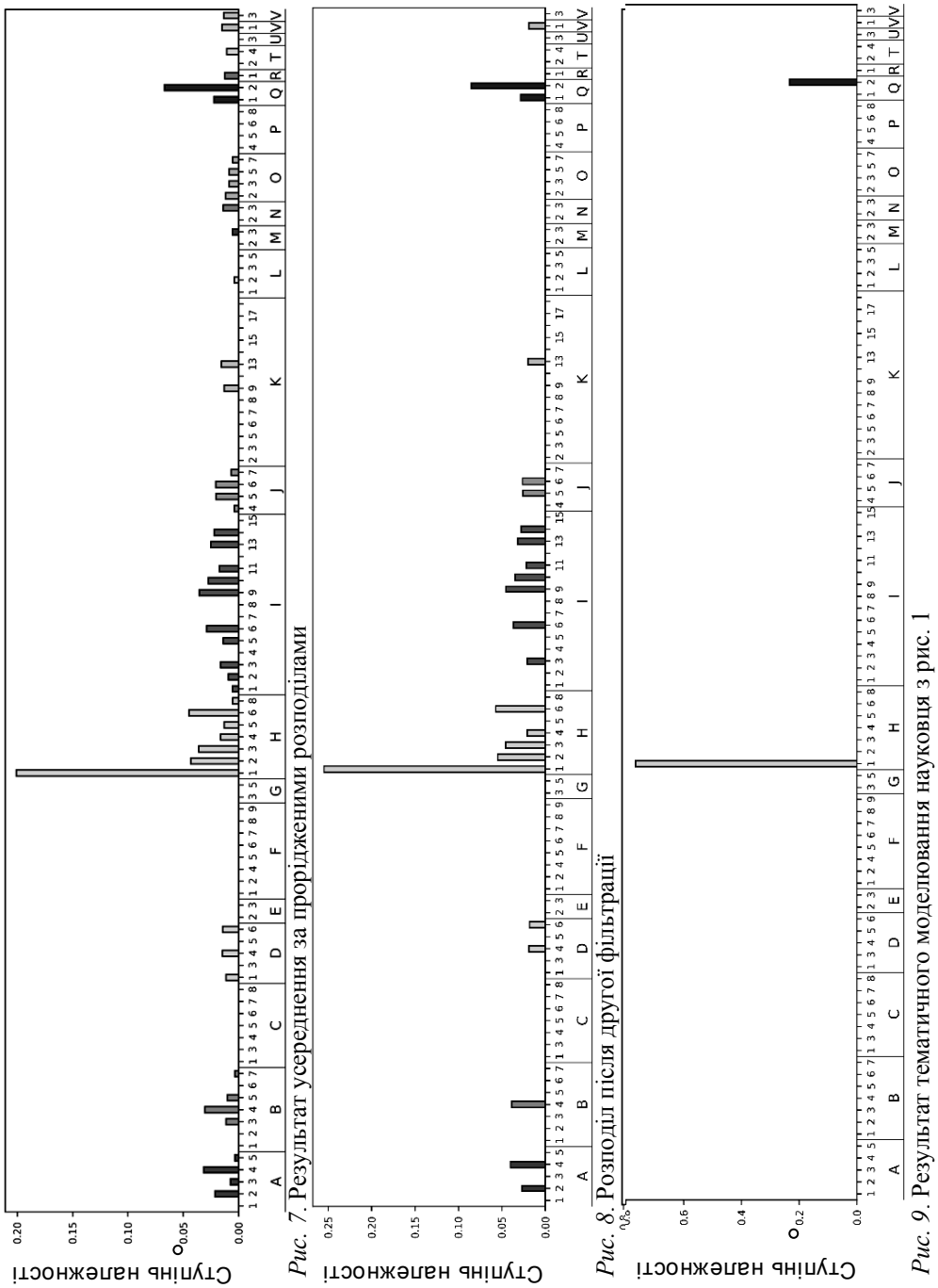

Рис. 7. Результат усереднення за проріджеными розподілами

Рис. 8. Розподіл після другої фільтрації

Рис. 9. Результат тематичного моделювання науковця з рис. 1

Yevgeniy Bodyanskiy

Kharkiv National University of Radio Electronics, Artificial Intelligence Department, Control

Підтверджено адресу електронної пошти в домені ikd.kiev.ua

Computational Intelligence Data Mining Data Stream Mining Big Data

Рис. 10. Профіль другого науковця

Nataliia Kussul (Наталия Куссуль)

Space research institute, National academy of science of Ukraine, Kiev

Підтверджено адресу електронної пошти в домені ikd.kiev.ua

Machine learning remote sensing data science
disaster management agricultural monitoring

Рис. 11. Профіль третього науковця

Порівнюючи результати, бачимо що за інтересами у Google Scholar, тобто за суб'єктивною інформацію дуже обмеженого обсягу, запропонований алгоритм достатньо добре категоризує науковців. Для кількісної оцінки узгодженості результатів скористаємося метрикою Чекановського. Для розглядуваного випадку — за умови нормованості суми належностей на 1, метрика Чекановського між двома науковцями W_1 і W_2 розраховується таким чином:

$$Fit(W_1, W_2) = \sum_{p=\overline{1, M}} \min(\mu_{t_p}(W_1), \mu_{t_p}(W_2)). \qquad (1)$$

Таблиця 2. Результати тематичного моделювання науковців

Спеціальність	Chernodub		Kussul		Bodyanskiy	
	Dimensions	Google Scholar	Dimensions	Google Scholar	Dimensions	Google Scholar
D6				0,283		
I9			0,675	0,447		
H1	0,8	0,767	0,172	0,346	0,797	0,295
H2						0,199
H6			0,153		0,203	0,506
K9	0,2					
Q2		0,233				

Метрику (1) можна інтерпретувати як суму ступенів належності перетину нечітких множин $\widetilde{W}_1$ і $\widetilde{W}_2$, які являють собою результати тематичного моделювання науковця за двома джерелами початкової інформації — за інтересами в Google Scholar та за категоризованими публікаціями в Dimensions.

За даними з табл. 2 отримуємо такі значення метрики (1):

Fit(Chernodub) = 0,767;
Fit(Bodyanskiy) = 0,498;
Fit(Kussul) = 0,619.

За метрикою (1) збіг враховується ізольовано — лише в межах кожної окремої спеціальності. Для врахування внеску споріднених спеціальностей пропонується до значення метрики (1) додати такий доданок:

$$\Delta Fit(W_1, W_2) = \sum_{v=\overline{1, M}} \sum_{p=\overline{1, M}} J(t_v, t_p) \cdot \min(\varepsilon_{t_v}(W_1), \varepsilon_{t_p}(W_2)), \qquad (2)$$

де $J(t_v, t_p)$ — індекс Жакара між спеціальностями t_v і t_p; $\varepsilon_{t_v}(W_1) = \max(0, \mu_{t_v}(W_1) - \mu_{t_v}(W_2))$ — залишок ступеня належності науковця до спеціальності t_v у $\widetilde{W}_1$ після врахування у формулі (1) збігу $\mu_{t_v}(W_1)$ і $\mu_{t_v}(W_2)$; $\varepsilon_{t_p}(W_2) = \max(0, \mu_{t_p}(W_2) - \mu_{t_p}(W_1))$ — залишок ступеня належності науковця до спеціальності t_p у $\widetilde{W}_2$ після врахування у формулі (1) збігу $\mu_{t_p}(W_1)$ і $\mu_{t_p}(W_2)$.

Для фільтрації інформаційного шуму формулу (2) застосуємо лише для пар спеціальностей з високою подібністю — з індексом Жакара понад 0,02. Для наведених в табл. 2 спеціальностей таких пар виявилося 3. Індекси Жакара для них є такими:

$$J(\text{D6, I9}) = 0{,}083;$$

$$J(\text{Hl, H6}) = 0{,}071;$$

$$J(\text{K9, Q2}) = 0{,}041.$$

Підставляючи числові дані у формулу (2), отримуємо:

$$\Delta Fit(\text{Chernodub}) = 0{,}008;$$

$$\Delta Fit(\text{Bodyanskiy}) = 0{,}022;$$

$$\Delta Fit(\text{Kussul}) = 0{,}03.$$

З урахуванням спорідненості спеціальностей збіг результатів тематичного моделювання трохи підвищився і становить:

$$Fit_{\text{sim}}(\text{Chernodub}) = 0{,}767 + 0{,}008 = 0{,}775;$$

$$Fit_{\text{sim}}(\text{Bodyanskiy}) = 0{,}498 + 0{,}022 = 0{,}52;$$

$$Fit_{\text{sim}}(\text{Kussul}) = 0{,}619 + 0{,}03 = 0{,}649.$$

ВИСНОВКИ

Запропоновано тематичне моделювання науковців на основі їх інтересів у профілях Google Scholar. Інтереси у профілях науковці вказують на власний розсуд без використання будь-якого словника ключових слів. Запропоновано підхід до категоризації таких науковців у межах системи класифікації наук ANZSRC. Відображення «науковець – спеціальності» здійснюється з використанням ресурсів інформаційної системи Dimensions, яка містить понад 110 млн наукових публікацій, що категоризовані за ANZSRC.

Алгоритм тематичного моделювання науковців містить три ділянки. На першій ділянці формується множина запитів на основі ключових слів та їх поєднань, на другій — відбувається тематичне моделювання за кожним запитом окремо з фільтрацією стоп-слів та маловживаних слів, а на третій — усереднюються належності за всіма запитами та обрізається розподіл до кількох спеціальностей. Під час вилучення мінорних спеціальностей враховується їх вплив на споріднені спеціальності. На виході алгоритму отри-

муємо ступені належності науковця до кількох спеціальностей, яким найбільше відповідає сукупність його інтересів. Таке відображення інтересів можна розглядати як аналог процедури word2vec.

Проведено порівняння тематичного моделювання на основі обмеженої інформації з профілів науковців з Google Scholar та за кількома десятками авторських статей, які категоризовано системою Dimensions. У результаті перевірки встановлено узгодженість результатів тематичного моделювання на основі різного обсягу початкової інформації. Це дозволяє використовувати запропонований алгоритм як основу технології інформаційної розвідки наукових кадрів, зокрема, для первинного підбору кандидатів у опоненти дисертацій, у рецензенти наукових проєктів для формування команди для виконання спільних наукових проєктів.

ЛІТЕРАТУРА

1. E. Delgado López-Cózar, E. Orduña-Malea, A. Martín-Martín, and J.M. Ayllón, "Google Scholar: the big data bibliographic tool", in *Research analytics: boosting university productivity and competitiveness through scientometrics*. CRC Press (Taylor & Francis), pp. 59–80, 2017. doi: 10.1201/9781315155890-4.
2. A. Martín-Martín, M. Thelwall, E. Orduna-Malea, and E.D. López-Cózar, "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations", *Scientometrics*, 126, pp. 871–906, 2021. doi: 10.1007/s11192-020-03690-4.
3. A.-W. Harzing and S. Alakangas, "Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison", *Scientometrics*, 106(2), pp. 787–804, 2016. doi: 10.1007/s11192-015-1798-9.
4. B. Rahdari et al., "Grapevine: A profile-based exploratory search and recommendation system for finding research advisors", *Proceedings of the Association for Information Science and Technology*, 57(1), e271, 2020. doi: 10.1002/pra2.271.
5. J. Saad-Falcon, O. Shaikh, Z.J. Wang, A.P. Wright, S. Richardson, and D.H. Chau, "PeopleMap: Visualization Tool for Mapping Out Researchers using Natural Language Processing", *arXiv preprint*, arXiv:2006.06105 (2020).
6. M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smith, "The author-topic model for authors and documents", in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press*, pp. 487–494, 2004.
7. D. Blei, A. Ng., and M. Jordan, "Latent Birichlet allocation", *Journal of Machine Learning Research*, 3, pp. 993–1022, 2003.
8. J. Jian, G. Qian, M. Haikun, and C. Chong, "Author–Subject–Topic model for Reviewer Recommendation", *JIS-Journal of Information Science, SAGE*, pp. 1–16, 2018. doi: 10.1177/0165551518806116.
9. D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers", in *KDD'07 proceedings of the 13th ACMSIGKDD international conference on knowledge discovery and data mining, New York: ACM*, pp. 500–509, 2007. doi: 10.1145/1281192.1281247.
10. N. Kawamae, "Author interest topic model", in *SIGIR'10 proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, New York: ACM*, pp. 887–888, 2010. doi: 10.1145/1835449.1835666.
11. C. Sun, T.J. King, P. Henville, and R. Marchant, "Hierarchical Word Mover Distance for Collaboration Recommender System", *Australasian Conference on Data Mining. Communications in Computer and Information Science, Springer 996*, pp. 289–302, 2018. doi: 10.1007/978-981-13-6661-1_23.

12. K. Xiangjie, J. Huizhen, Y. Zhuo, Y. Zhuo, Y. Zhuo, and A. Tolba, "Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation", *PlosOne*, 11(2), e0148492, 2016. doi: 10.1371/journal.pone.0148492
13. Y. Zhao, J. Tang, and Z. Du, "EFCNN: A Restricted Convolutional Neural Network for Expert Finding", in *Advances in Knowledge Discovery and Data Mining. PAKDD 2019. Lecture Notes in Computer Science*, vol. 11440, Springer, Cham, 2019. doi: 10.1007/978-3-030-16145-3_8.
14. A. Omer, G. Hongyu, B. Suma, H. Wen-Mei, and X. JinJun, "PaRe: A Paper Reviewer Matching Approach Using a Common Topic Space", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP)*, pp. 518–528, 2019. doi: 10.18653/v1/D19-1049.
15. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", in *Proceedings of the 26th International Conference on Neural Information Processing Systems 2*, pp. 3111–3119, 2013.
16. T. Hofmann, "Probabilistic latent semantic indexing", in *Proc. 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, 1999. doi: 10.1145/312624.312649.
17. S. Shtovba and M. Petrychko, "Jaccard Index-Based Assessing the Similarity of Research Fields in Dimensions", *CEUR Workshop Proceedings*, vol. 2533 "Proc. of the First International Workshop on Digital Content & Smart Multimedia", pp. 117–128, 2019.

Надійшла 17.03.2021

INFORMATION ON THE ARTICLE

Serhiy D. Shtovba, ORCID: 0000-0003-1302-4899, Vasyl Stus' Donetsk National University, Vinnytsia, Ukraine, e-mail: s.shtovba@donnu.edu.ua

Mykola V. Petrychko, ORCID: 0000-0001-6836-7843, Vinnytsia National Technical University, Vinnytsia, Ukraine, e-mail: mpetrychko@vntu.edu.ua

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ УЧЕНЫХ НА ОСНОВЕ ИХ ИНТЕРЕСОВ В GOOGLE SCHOLAR / С.Д. Штовба, Н.В. Петрычко

Аннотация. Предложен алгоритм тематического моделирования ученых по научным специальностям на основе их интересов в профилях в Google Scholar. Алгоритм использует перечень научных специальностей из системы классификации наук ANZSRC. Информационным ресурсом для тематического моделирования является база категоризированных научных публикаций из системы Dimensions. Интересы из профилей ученых используются как поисковые запросы для Dimensions, сервисы которой выдают распределения релевантных документов по специальностям. Для уменьшения информационного шума эти распределения проходят несколько этапов обработки. Сравниваются результаты тематического моделирования на основе профильных интересов в Google Scholar и категоризированного списка авторских публикаций в Dimensions по метрике Чекановского с учетом схожести специальностей. Для тестовых ученых выявлена высокая согласованность результатов тематического моделирования при различной исходной информации.

Ключевые слова: тематическое моделирование, категоризация, Google Scholar, Dimensions, ANZSRC, профиль ученого, научные интересы, метрика Чекановского, индекс Жакарра.

TOPIC MODELING OF RESEARCHERS BASED ON THEIR INTERESTS FROM GOOGLE SCHOLAR / S.D. Shtovba, M.V. Petrychko

Abstract. The article proposes an algorithm for topic modeling of researchers based on their interests from Google Scholar profiles. The algorithm uses the set of fields of research from research classification system ANZSRC. An information resource for topic modeling is a corpus of categorized publications from Dimensions. Interests from researchers' profiles are used as search queries to Dimensions that outputs distributions of documents over categories. To reduce information noise these distributions are taken through a few stages of processing. The article also compares the results of topic modeling based on interests from Google Scholar profiles and based on a categorized list of publications from Dimensions. The comparison is done using modified Czekanowski metric that takes into account the similarity between categories. The results of comparing the topic modeling outputs based on different information sources show a good match.

Keywords: topic modeling, categorization, Google Scholar, Dimensions, ANZSRC, researcher's profile, research interests, Czekanowski metric, Jaccard index.