УДК 001.2

С. Д. Штовба
М. В. Петричко
Р. О. Тилець

Статистичний підхід до оцінювання подібності наукових спеціальностей в системі Dimensions

Вінницький національний технічний університет

Анотація
Запропоновано метод оцінювання подібності наукових спеціальностей. Метод може застосовуватися для створення рекомендаційних систем із підбору науковців зі спорідненими тематиками досліджень. Метод реалізовано для переліку наукових спеціальностей в системі Australian and New Zealand Standard Research Classification. Для статистичної оцінки коефіцієнтів подібності використана база публікацій Dimensions.
Ключові слова: класифікація наук, наукометрія, споріднені спеціальності, Dimensions, ANZSRC.

Abstract
A method for estimating the similarity of scientific fields is proposed. The method can be used to create recommendation systems for selecting scholars with mutual research topics. The method is fitted on the Australian and New Zealand Standard Research Classification system. Statistical estimation of similarity coefficients is based on Dimensions publications base.
Keywords: classification of sciences, scientometrics, neighboring fields, Dimensions, ANZSRC.

З швидким розвитком інформаційних технологій зростає зв'язність між учасниками різноманітних спільнот, в тому числі і наукових. Постає складна задача кластеризації учасників спільнот на групи з схожими інтересами. Стосовно наукових спільнот – це виявлення груп науковців, що працюють в схожих напрямках. Актуальність таких задач в Україні стрімко зросла в зв'язку із запровадженням в березні 2019р. експерименту щодо захисту PhD-дисертацій в одноразових спецрадах. Однією із ключових умов формування одноразових спецрад є близькість наукового напрямку дисертації та наукових досліджень членів спецради. Виникає питання як формалізувати цю відстань, щоб автоматично верифікувати відповідність складу спецради науковому напрямку дисертації. Аналогічна задача виникає в рекомендаційних системах із пошуку партнерів для спільних наукових досліджень, пошуку друзів у наукових соціальних мережах тощо. Доведено, що співпраця між науковцями має високий вплив на наукову продуктивність [1]. Відповідно, реалізуються технології підбору партнерів, які зазвичай використовують бібліографічну інформацію з публікацій [2].

Окрім бази публікацій, є інформація про науковців більш високого ієрархічного рівня, наприклад, тематика його досліджень. Ця тематика може бути як формалізованою, так і неформалізованою. Приклад неформалізованої тематики – це ключові слова публікацій або список інтересів в профілі науковця в Google Scholar. Приклад формалізованої тематики – наукові спеціальності дослідника за певною системою класифікації наук. В попередніх дослідженнях [3] ми розробили інформаційну технологію категоризації науковців. Вона полягає у переході від неформалізованої тематики у формалізовану. Зокрема, здійснюється відображення інтересів науковця з профіля в Google Scholar в спеціальності та галузі наук за системою ANZSRC – Australian and New Zealand Standard Research Classification. Приклад такої категоризації наведено на рис. 1. З цього рисунка видно, що науковець може належати до різних спеціальностей, але з

різним ступенем. Виникає задача як розрахувати коефіцієнт подібності, який покаже ступінь схожості двох науковців.

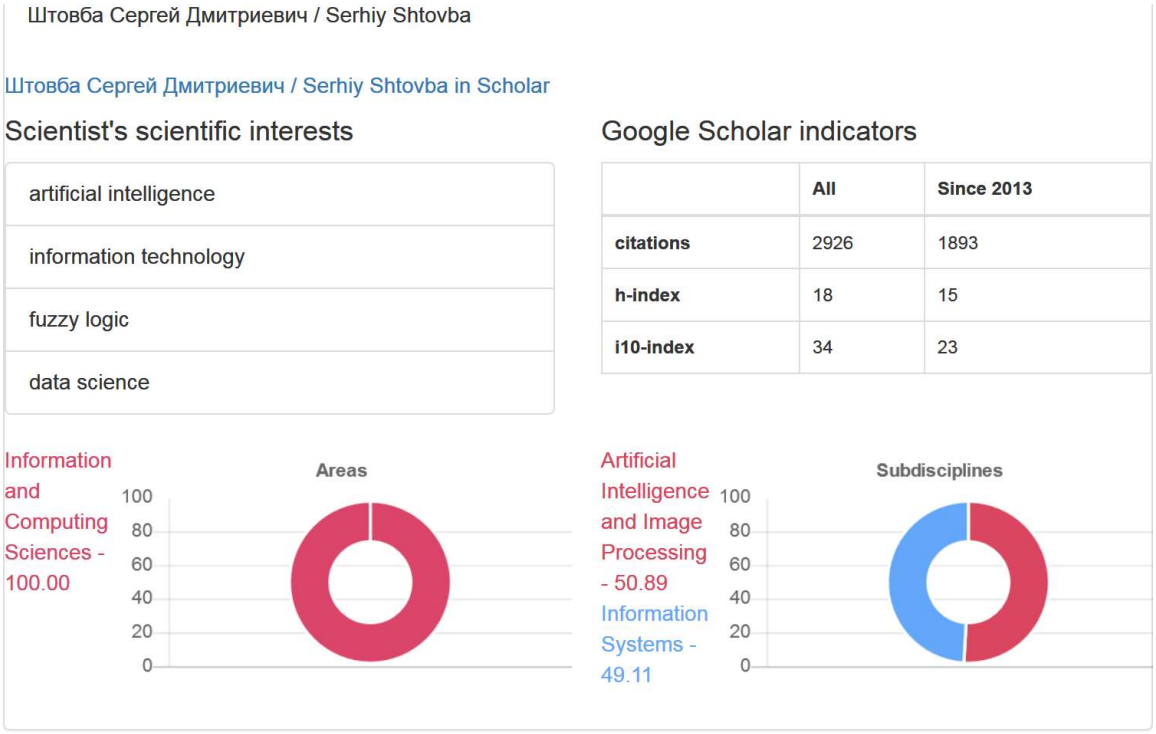

Рис. 1. Категоризація науковця за спеціальностями з ANZSRC

Для визначення подібності можна застосовувати підходи з екології та ботаніки, наприклад, коефіцієнти Жакарда та Чекановського. Але вони враховують подібність за бінарним принципом. Стосовно науковців подібність визначається через агрегування перетину за кожною спеціальністю. При цьому не враховується внесок споріднених спеціальностей, наприклад, геології та геохімії. Метою нашого дослідження, є чисельна оцінка подібності споріднених спеціальностей. Оцінку проведено для переліку спеціальностей ANZSRC з використанням інформаційної системи Dimensions.

Dimensions – частково безкоштовна наукова база даних запущена компанією Digital Science у січні 2018 року. Вона індексує біля 100 млн публікацій. Для впорядкованості публікацій використовується спрощений варіант ANZSRC, в якому наука поділена на 22 галузі та 154 спеціальності. Для своїх експериментів ми будемо використовувати її як джерело інформації.

Ми пропонуємо метод визначення подібності наукових спеціальностей, що базується на кількості публікацій які віднесені до цих спеціальностей. Для визначення коефіцієнту подібності введемо наступні позначення:

$S1$ – перша спеціальність;

$S2$ – друга спеціальність;

$k1$ – кількість документів по першій спеціальності;

$k2$– кількість документів по другій спеціальності;

C – кількість документів з галузями 1 та 2.

Тоді подібність цих двох галузей визначатиметься як:

$$similarity(S1, S2) = \frac{C}{k1 + k2}$$

Значення $similarity(S1, S2)$ знаходиться в діапазоні $[0,1]$, де 0 означає відсутність подібності, 1 – ідентичність.

Для усіх пар спеціальностей можна сформувати матрицю коефіцієнтів подібності:

$$similarity(S_i, S_j) = \frac{C_{ij}}{k_i + k_j}$$

де $i = \overline{1,n}$, $j = \overline{1,n}$, n – кількість спеціальностей; C_{ij} – кількість документів, які відносяться одночасно до i-ї та j-ї спеціальності; k_i – кількість документів, що віднесені до i-ї спеціальності; k_j - кількість документів, що віднесені до j-ї спеціальності.

Розглянемо приклад визначення коефіцієнту подібності використовуючи систему Dimensions. Візьмемо дві спеціальності Geology та Geochemistry. Дана система надає статистику кількості публікацій по кожній спеціальності науки.

Таблиця 1 – Показники подібності спеціальностей Geology та Geochemistry

Спеціальність	Кількість публікацій	Кількість спільних публікацій одночасно віднесених до Geology та Geochemistry	Коефіцієнт подібност
Geology	185265	25994	0.1059
Geochemistry	60146		

Маючи таку статистику можна визначити подібність між будь-якими двома спеціальностями у даній системі наук. У таблиці 2 подано 30 найбільш подібних спеціальностей згідно із використаною системою наук та статистикою на основі Dimensions.

Таблиця 2 – Найбільш подібні наукові спеціальностей в ANZSRC / Dimensions

Спеціальність 1	Спеціальність 2	Подібність
Specialist Studies In Education	Curriculum and Pedagogy	0,211131387
Ecology	Environmental Science and Management	0,188713844
Geology	Geochemistry	0,10607919
Applied Economics	Econometrics	0,087915677
Plant Biology	Crop and Pasture Production	0,082960643
Linguistics	Cognitive Sciences	0,070854084
Historical Studies	Political Science	0,070221075
Physical Chemistry (incl. Structural)	Materials Engineering	0,069963051
Cultural Studies	Other Studies In Human Society	0,065344136
Linguistics	Language Studies	0,064295204
Banking, Finance and Investment	Econometrics	0,064150666
Ecological Applications	Other Biological Sciences	0,063932623
Artificial Intelligence and Image Processing	Information Systems	0,059087625
Sociology	Policy and Administration	0,058747766
Historical Studies	Literary Studies	0,058470565
Information Systems	Computer Software	0,057940991
Forestry Sciences	Ecological Applications	0,057695828
Applied Economics	Banking, Finance and Investment	0,055657539
Biochemistry and Cell Biology	Genetics	0,05523122
Physical Chemistry (incl. Structural)	Macromolecular and Materials Chemistry	0,05462369
Historical Studies	Cultural Studies	0,054268764
Political Science	Policy and Administration	0,053932274
Environmental Science and Management	Ecological Applications	0,052267514
Econometrics	Economic Theory	0,04911794
Geology	Geophysics	0,047260966
Computer Software	Computation Theory and Mathematics	0,046940267
Geology	Physical Geography and Environmental Geoscience	0,045209256
Cultural Studies	Literary Studies	0,044653537
Public Health and Health Services	Psychology	0,043305142
Ecology	Ecological Applications	0,042745563

З таблиці 2 видно, що коефіцієнт подібності дуже різко зменшується. Це свідчить про можливу наявність шумових значень. Для відкидання шумових значень можна використати інтегральний розподіл. Для цього визначається кумулятивна сума коефіцієнтів подібності. Інтегральна крива розподілу коефіцієнтів подібності подана на рисунку 2.

Рис. 2 Інтегральна крива коефіцієнтів подібності
(по осі X – номери впорядкованих пар спеціальностей,
по осі Y – коефіцієнт подібності відповідної пари)

З інтегрального розподілу видно, що приблизно 100 пар спеціальностей вносять 51% подібності. Усі решта мають коефіцієнт подібності менше 2%. Таким чином з усіх комбінацій спеціальностей лише 100 пар мають вагомий коефіцієнт подібності. Усі інші можна вважати не схожими.

Одне із застосувань коефіцієнту подібності спеціальностей є пошук найближчих науковців. Припустимо, що у нас є база науковців, що віднесені з деяким ступенем належності до деякої спеціальності/спеціальностей із заданої системи наук, тоді найпростіший коефіцієнт подібності науковців має вигляд:

$$similarity(r1, r2) = sum(\min(A, B)),$$

де $r1, r2$ – деякі науковці, A та B – вектори ступенів належності до спеціальності/спеціальностей відповідних науковців. Така міра подібності називається ще мірою Чекановського. Вона має декілька недоліків. Один з них це те, що вона не враховує споріднені спеціальності. Очевидно, що якщо науковці віднесені до споріднених спеціальностей за даною ознакою подібності подібність дорівнюватиме нулю, оскільки їх перетин (мінімум) це нуль. Дану ознаку можна вдосконалити записавши:

$$similarity(r1, r2) = \sum_{j=0}^{n} \min(A_j, B_j) + \sum_{i=0}^{n} \sum_{j=0}^{n} similarity(s_i, s_j) * \min(A_i, B_j),$$

де $i = \overline{1,n}$, $j = \overline{1,n}$, n – кількість спеціальностей; A_i – ступінь належності науковця $r1$ до i-ї спеціальності, B_j – ступінь належності науковця $r2$ до j – спеціальності; s_i – i-та спеціальність, s_j – j-та спеціальність. Таким чином, за рахунок другої складової ознаки подібності два науковці з спорідненими спеціальностями завжди матимуть деяке значення подібності.

Розглянемо приклад знаходження подібності двох науковців. Для цього необхідно мати науковців, що віднесені до певної спеціальності/спеціальностей науки. Використаємо науковців з нашої системи категоризації науковців [3]. Для прикладу візьмемо двох науковців: Штовбу С.Д. та Бісікала О.В.. Порівняємо їхню схожість за подібністю спеціальностей науки. Розрахунки зведемо в табл. 3. Таким чином на основі подібності спеціальностей науки нам вдалось збільшити подібність науковців зі спорідненими спеціальностями науки.

Таблиця 3 – До розрахунку подібності наукових напрямків двох науковців

Науковець	Ступінь належності до спеціальності	Схожість Artificial Intelligence and Image Processing та Cognitive Sciences	Схожість Artificial Intelligence and Image Processing та Linguistics	Схожість Information Systems та Linguistics	Схожість Information Systems та Cognitive Sciences	Схожість Artificial Intelligence and Image Processing та Information Systems	Перший показник подібності	Другий показник подібності (на основі подібності спеціальностей науки)
Бісікало О.В.	Artificial Intelligence and Image Processing – 0.396 Linguistics – 0.316 Cognitive Sciences – 0.286	0,0037184724	0,003836850	0,00269967	0,00086764	0,059087624	0.396	0,4179330
Штовба С.Д.	Artificial Intelligence and Image Processing – 0.508 Infromation Systems – 0.491							

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1.Lopes GR, Moro MM, Wives LK, De Oliveira JPM (2010) Collaboration recommendation on academic social networks. In: Advances in Conceptual Modeling–Applications and Challenges, Springer. pp. 190–199.

2. Xiangjie Kong, Huizhen Jiang, Zhuo Yang, Zhuo Yang, Zhuo Yang (2016) Amr Tolba Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation – PlosOne.

3. Штовба С.Д., Петричко М.В. Автоматична категоризація науковців за тематикою досліджень на основі профілей в Google Scholar / С.Д. Штовба, М.В. Петричко / Матеріали XLVII Наук.-техн. конф. факультету КСА ВНТУ, Вінниця, 21-23 березня 2018 р. https://conferences.vntu.edu.ua/public/files/1/fksa_2018_netpub.pdf. – С. 1561-1578.

Сергій Дмитрович Штовба *– д.т.н., професор кафедри комп'ютерних систем управління, Вінницький національний технічний університет, м. Вінниця, e-mail: shtovba@vntu.edu.ua.*

Микола Володимирович Петричко *– ст.гр. 2АКІТ-18м, факультету комп'ютерних систем та автоматики Вінницького національного технічного університету, м. Вінниця, e-mail: petrychko.myckola@gmail.com.*

Тилець Роман Олексійович*, аспірант кафедри комп'ютерних систем управління, Вінницький національний технічний університет, м. Вінниця, e-mail: кафедри комп'ютерних систем управління, Вінницький національний технічний університет, м. Вінниця, e-mail: shtovba@vntu.edu.ua.*

Shtovba Serhiy *—Professor on Department of Computer Control Systems, Vinnytsia National Technical University, Vinnytsia, e-mail: shtovba@vntu.edu.ua.*

Petrychko Mykola *— student, Facultyt of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, email : petrychko.myckola@gmail.com.*

Tylets Roman *– PhD-student on Department of Computer Control Systems, Vinnytsia National Technical University, Vinnytsia, e-mail:, Vinnytsia, e-mail: roman.tylets@gmail.com.*