УДК 001.2

С. Д. Штовба
М. В. Петричко

Модифікована метрика Чекановського для оцінки схожості науковців з врахуванням спорідненості спеціальностей

Вінницький національний технічний університет

Анотація
Запропоновано метрику для оцінки схожості науковців на основі метрики Чекановського у випадку представлення інтересів науковців у вигляді нечітких множин. Носієм нечіткої множини є спеціальності науки визначеної згідно з стандартом Australian and New Zealand Standard Research Classification (ANZSRC). При порівнянні двох нечітких множин пропонується враховувати також попарну схожість наукових спеціальностей для досягнення вищого рівня схожості.
Ключові слова: Google Scholar, профіль науковця, науковець, метрика Чекановського, *ANZSRC, наукові спеціальності, індекс Жакара.*

Abstract
A metric for assessing the similarity between researchers based on Chekanovskiy's metric with researchers represented as fuzzy sets is proposed. Support of a fuzzy sets are research specialties defined by Australian and New Zealand Standard Research Classification (ANZSRC). When comparing two fuzzy sets it is proposed to take into account also a pairwise similarity between research specialties to get a higher level of similarity.
Keywords: Google Scholar, researcher's profile, researcher, Czekanowski metric, ANZSRC, research specialties, Jaccard index.

Задача порівняння науковців має важливе місце у таких системах, де виконується пошук науковців для спільних досліджень. Роботи над цією тематикою мали місце у [1, 2]. Одне з інших застосувань це пошук експертів за деякими ключовими словами (ключові слова представляються у тому ж вигляді що й профіль науковця) [3]. Для рекомендації рецензентів [4, 5] на рецензовані роботи також може застосовуватись порівняння науковців. Зазвичай у таких випадках порівнюється профіль науковця із рецензованою робою. І профіль науковця, і рецензована робота представляються у вигляді деякого вектору, що нічим не відрізняється від порівняння двох профілів науковців.

У згаданих системах проблема порівняння зводиться до проблеми представлення даних про науковців у вигляді деяких векторів, що відображають його діяльність та/або наукові інтереси. Сформувавши такі вектори задача порівняння науковців зводиться до задачі порівняння векторів чисел, що є досить формалізованою. Існує багато способів представити профіль науковця у вигляді деякого вектору чи векторів. Деякі представлення на основі статистики були приведені у [6, 7], у яких по публікаціях науковця будувався розподіл по деякому словникові слів. У даній роботі за основу представлення науковців обрано спосіб поданий у роботі [8], де науковець представляється у вигляді нечіткої множини на універсальні множині наукових спеціальностей з системи ANZSRC.

Традиційно для порівняння векторів застосовується косинусоїдна метрика. У цій роботі для науковців представлених у вигляді нечітких множин пропонується використовувати метрику Чекановського, оскільки її формулювання має кращу інтерпретацію. Для науковців з інтересами W_1 та W_2 метрика Чекановського записується наступним чином:

$$Fit(W_1, W_2) = \sum_{p=\overline{1,m}} \min\left(\mu_{t_p}(W_1), \mu_{t_p}(W_2)\right) \quad (1)$$

де $\mu_{t_p}(W_2)$ - ступінь належності сукупності інтересів W_1 до спеціальності t_p, $\mu_{t_p}(W_2)$ - ступінь належності сукупності інтересів W_2 до спеціальності t_p, $p = \overline{1,m}$. Операція мінімуму є перетином нечітких множин.

Для прикладу розглянемо результат порівняння науковців з Google Scholar поданих на рисунку 1. Їх представлення у вигляді нечітких множин подано у таблиці 1. За цими даними отримуємо наступні значення метрики схожості Чекановського:

$$Fit(W_{kussul}, W_{bodyanskiy}) = 0.295$$

Таблиця 1 – Представлення науковців у вигляді нечітких множин

Код спеціальності	Kussul	Bodyanskiy
0406	0.283	
0901	0.447	
0801	0.346	0.295
0802		0.199
0806		0.506

Yevgeniy Bodyanskiy

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

Подтвержден адрес электронной почты в домене nure.ua

Computational Intelligence Data Mining Data Stream Mining Big Data

Nataliia Kussul (Наталия Куссуль)

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

Подтвержден адрес электронной почты в домене ikd.kiev.ua

Machine learning remote sensing data science disaster management agricultural monitoring

Рис. 1. Тестові профілі науковців для порівняння

За метрикою Чекановського схожість рахується попарно між спільними науковими спеціальностями. При цьому не враховується перехресна схожість адже наукові спеціальності можуть бути частиною більш загальної науки. Таким чином, науковці, що займаються у схожих спеціальностях науки повинні мати більший рівень схожості. Для врахування цього пропонується у метрику Чекановського додати наступний доданок:

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

де $J(t_v, t_p)$ – індекс Жакара між спеціальностями t_v та t_p [7]; $\varepsilon_{t_v}(W_1) = \min(0, \mu_{t_v}(W_1) - \mu_{t_v}(W_2))$ – залишок ступеня належності науковця до спеціальності t_v в $\widetilde{W}_1$ після обліку в (1) співпадіння між $\mu_{t_v}(W_1)$ та $\mu_{t_v}(W_2)$; $\varepsilon_{t_p}(W_2) = \min(0, \mu_{t_p}(W_2) - \mu_{t_p}(W_1))$ – залишок ступеня належності науковця до спеціальності t_p в $\widetilde{W}_2$ після обліку в (1) співпадіння між $\mu_{t_p}(W_1)$ та $\mu_{t_p}(W_2)$ в (1).

Для фільтрації інформаційного шуму, формулу (2) застосуємо лише для пар спеціальностей з високою подібністю – з індексом Жакара понад 0.02. Для наведених в табл. 1 спеціальностей є лише одна така пара. Індекс Жакара для неї є такими:

$J(0801, 0806) = 0.071$

Підставляючи числові дані в (2), отримуємо:

$$Fit(W_{kussul}, W_{bodyanskiy}) = 0.295 + 0.071 = 0.366$$

Висновки

Запропонована модифікація метрики Чекановського для порівняння двох науковців представлених у вигляді нечітких множин, термами яких є наукові спеціальності, дозволяє збільшити схожість між науковцями шляхом визначення схожості між науковими спеціальностями.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Sun C., King T. J., Henville P., Marchant R. Hierarchical Word Mover Distance for Collaboration Recommender System. Australasian Conference on Data Mining. Communications in Computer and Information Science, Springer 996, 289-302 (2018). DOI: 10.1007/978-981-13-6661-1_23.
2. Xiangjie K., Huizhen J., Zhuo Y., Zhuo Y., Zhuo Y., Tolba A. Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation. PlosOne 11(2): e0148492 (2016). DOI: 10.1371/journal.pone.0148492
3. Zhao Y., Tang J., Du Z. EFCNN: A Restricted Convolutional Neural Network for Expert Finding. In: Yang Q., Zhou ZH., Gong Z., Zhang ML., Huang SJ. (eds) 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.
4. Omer A., Hongyu G., Suma B., Wen-Mei H., JinJun X.. 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), 518–528 (2019). DOI: 10.18653/v1/D19-1049.
5. Mimno D., McCallum A. Expertise modeling for matching papers with reviewers. In: KDD'07 proceedings of the 13th ACMSIGKDD international conference on knowledge discovery and data mining, 500–509 (2007). New York: ACM. DOI: 10.1145/1281192.1281247.
6. Rosen-Zvi M., Griffiths T., Steyvers M., Smith P. The author-topic model for authors and documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press, 487-494 (2004).
7. Jian J., Qian G., Haikun M., Chong C. Author–Subject–Topic model for Reviewer Recommendation. JIS-Journal of Information Science, SAGE, 1-16 (2018). DOI: 10.1177/0165551518806116.
8. 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.
9. Shtovba S., Petrychko M. 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". – 2019. – P. 117-128.

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 *– аспірант, факультету комп'ютерних систем та автоматики Вінницького національного технічного університету, м. Вінниця, e-mail: mpetrychko@vntu.edu.au.*

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

Petrychko Mykola *— PhD student, Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, email : mpetrychko@vntu.edu.ua.*