УДК 001.2

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

Автоматична категоризація науковців на основі профілей в Google Scholarз врахуванням спорідненості між науковими спеціальностями

Вінницький національний технічний університет

Анотація
Покращено алгоритм категоризації науковців на основі ключових слів шляхом розширення множини ключових слів їх попарним сполученням та використанням схожості наукових спеціальностей для збільшення впливу наукових спеціальностей, що сильно взаємодіють між собою.
Ключові слова: Google Scholar, ANZSRC, наукометрія, профіль науковця, наукові інтереси, категоризація науковців, рекомендаційні системи, схожість наукових спеціальностей.

Abstract
Improved the algorithm for researchers categorization based on keywords. Extended the set of keywords with their pairwise conjunction and used research specialties' similarity to increase research specialties' influence that are cooperating closely.
Keywords: Google Scholar, ANZSRC, scientometrics, researcher's profile, research interests, researchers' categorization, recommender systems, research specialties' similarity.

Останнім часом в задачах класифікації акценти змістили в бік обробки неструктурованої або слабкоструктурованої інформації, зокрема різноманітних природномовних текстів, які генеруються користувачами. Цей контент віддзеркалює уявлення користувача щодо того чи іншого явища, але це уявлення не формалізовано в межах деякої змістовної структури. Текстове вираження думок користувача значною мірою залежить від його персональних знань та лексикону. Відповідну одну і ту саму думку, люди висловлюють у різний спосіб. Така діяльність характерна для он-лайнових соціальних мереж.

Наукові спільноти теж об'єднані в різноманітні мережі. Найбільшою серед них є Google Scholar. В цій мережі у відкритому доступі є понад 50 тисяч профілів українських науковців. Науковець в профілі вказує свої інтереси, при чому робить він це на власний розсуд, обираючи слова у довільний спосіб. Google Scholar дозволяє здійснити пошук науковців за тим чи іншим інтересом. Але видачі формуються за буквальним співпадінням. Наприклад, видачі для "fuzzy set" та "fuzzy sets" будуть різними (рис. 1 та рис. 2), не говорячи вже синонімічні інтереси типу "fuzzy reasoning" та "fuzzy inference" (рис. 3 та рис. 4). Також, Google Scholar під час пошуку не враховує і сукупність інтересів користувача. Таким чином, пошукові та аналітичні сервіси за велетенським масивом профілів науковців в Google Scholar є досить примітивними.

Існують задачі з відбору релевантних осіб, вирішення яких спрощується за наявності категоризованих науковців. Одна з таких задач – підбір рецензента для написання відгуку. Роботи над цією тематикою мали місце у [1-5]. В роботі [5] запропонована рекомендаційна систему для підбору науковців зі спільними дослідженнями. Автори роботи поєднують тематичне моделювання, word2vec та так звану метрику mover distance для визначення подібності науковців на основі анотацій їх статей. За результатами система є успішною у визначенні існуючих спів-

авторств з вибірки даних не зважаючи на те, що ця інформація не були використана при моделюванні. І при цьому пропонує коректні потенційні можливості для співпраці із споріднених галузей дослідження незалежно від попередньої співпраці. Подібність науковців визначається на основі подібності їх публікацій, подібність публікацій визначається на основі подібності тематичного розподілу публікації, подібність тематичного розподілу визначається на основі подібності слів. За своєю суттю цей алгоритм представляє науковця у вигляді вектору, де кожний елемент це галузь науки (рис. 5). Звісно, він може не відображати реального стану оскільки формування цього вектору здійснюється на основі усіх статей науковця, що можуть бути застарілими та писатись кількома співавторами.

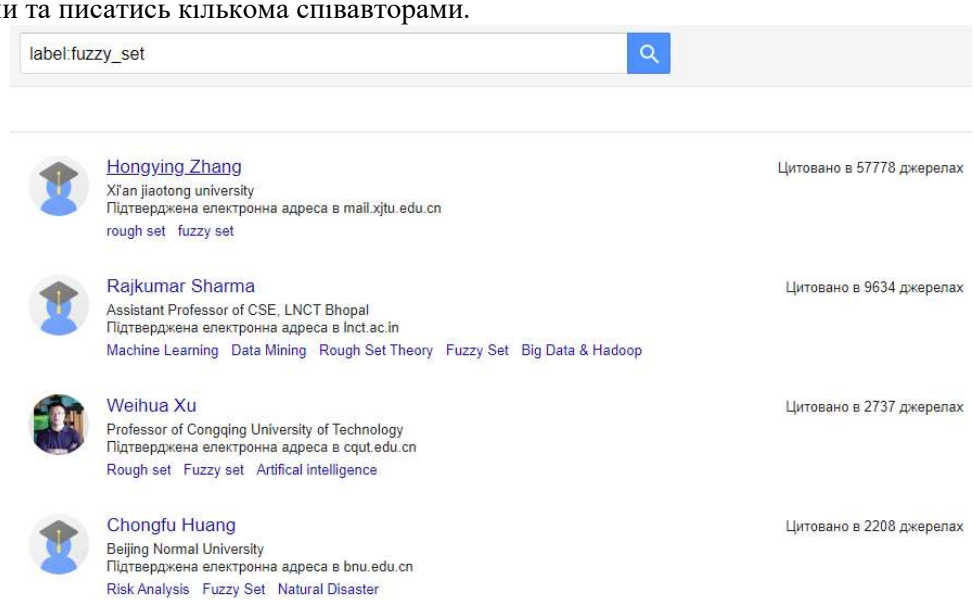

Рис. 1. Пошук профілів у Google Scholar за ключовим словом “fuzzy set”

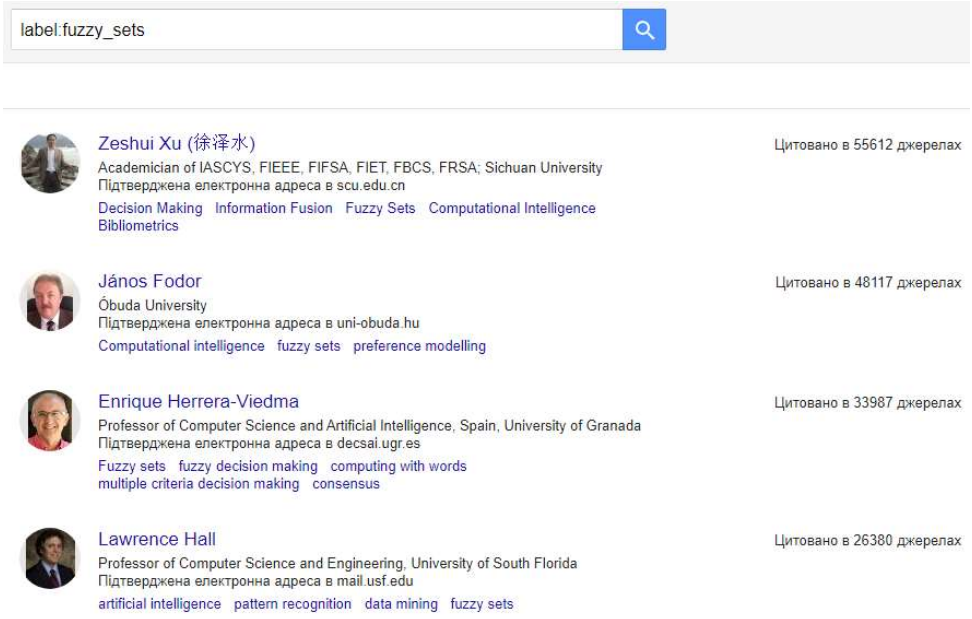

Рис. 2. Пошук профілів у Google Scholar за ключовим словом “fuzzy sets”

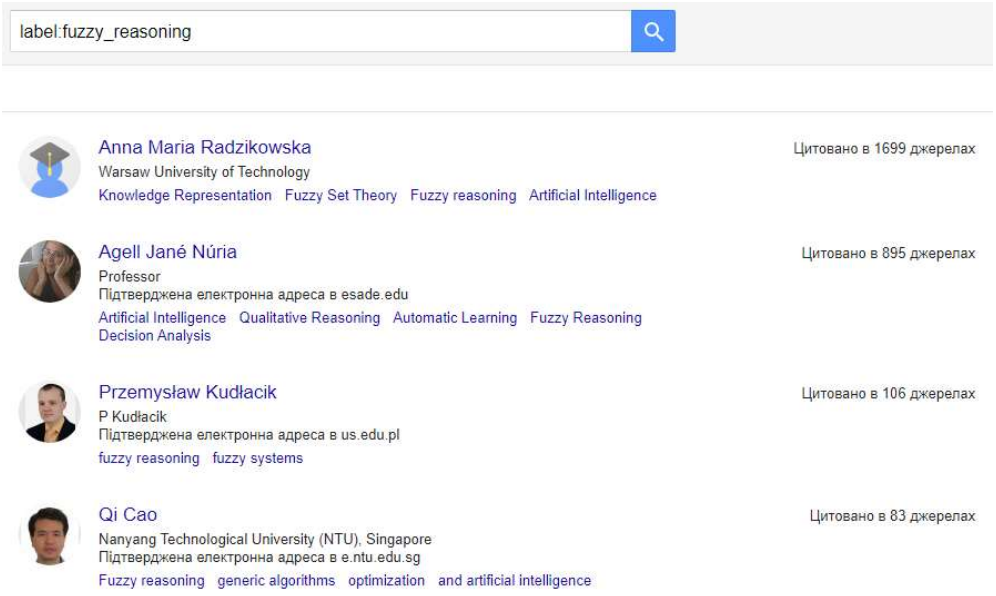

Рис. 3. Пошук профілів у Google Scholar за ключовим словом “fuzzy reasoning”

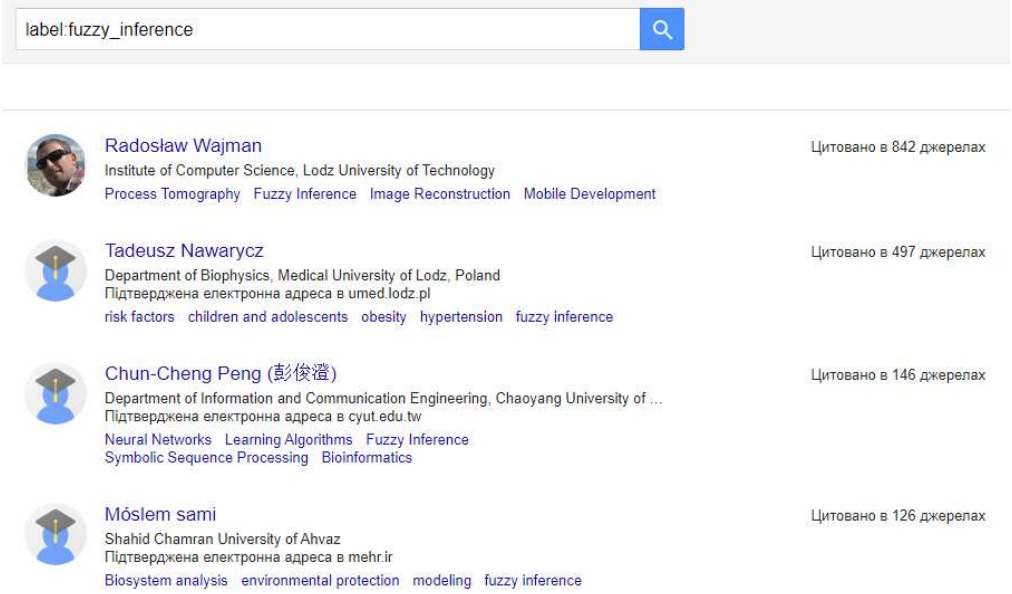

Рис. 4. Пошук профілів у Google Scholar за ключовим словом “fuzzy inference”

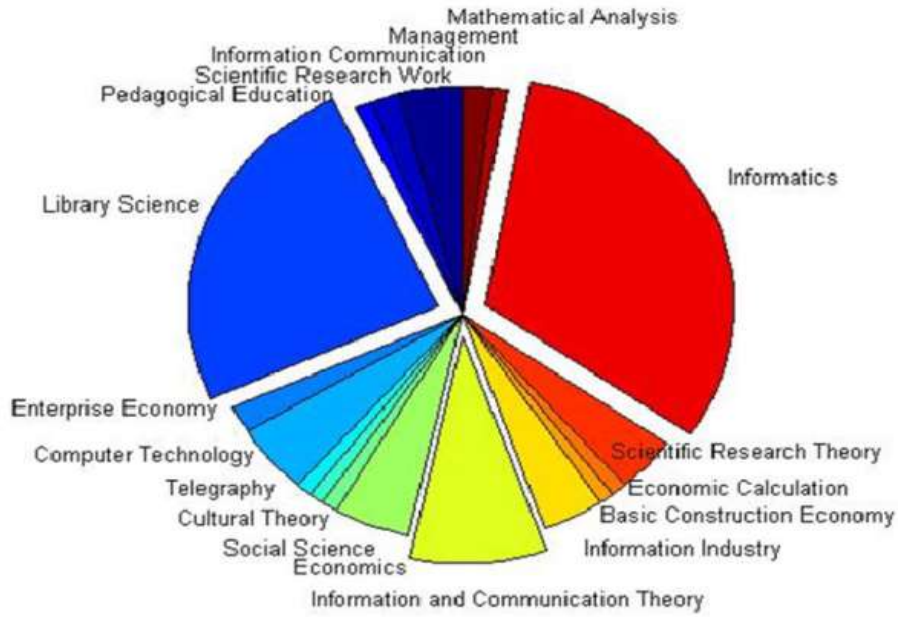

Рис. 5. Профіль науковця у вигляді розподілу по галузям науки [5]

Результати цієї роботи базуються на попередніх роботах авторів [6, 7] про автоматичну категоризацію на основі ключовий слів (наукових інтересів) та про визначення подібності між науковими спеціальностями. Використовуючи результати попередніх робіт ми покращили категоризацію науковців шляхом застосування парних ключових слів та подібності наукових спеціальностей. У якості системи класифікації наук використовується ANZSRC –Australian and New Zealand Standard Research Classification.

Для покращення категоризації множину ключових слів було розширено додаючи попарне сполучення ключових слів через роздільних "&". Наприклад, якщо у науковця множина ключових слів містить {"fuzzy logic", "medicine"}, то розширена множина виглядатиме ось так: {"fuzzy logic", "medicine", "fuzzy logic & medicine"}. Це дозволить враховувати не лише появу окремих ключових слів у базі розмічених документів, але і одночасну появу пари ключових слів, що дозволить підкреслити саме взаємодію цих ключових слів.

Мотивація для використання перерахунку ступенів належності на основі схожості спеціальностей пояснюється тим, що якщо у розподілі багато споріднених спеціальностей то має сенс їхній загальний внесок на результат категоризації. Формально перерахунок визначений у вигляді формули:

$$\bigvee_{i=0}^{n-1} A_i = A_i + A_i * similarity(S_i, S_n), \quad (1)$$

де A – множина ступенів належності, S – множина тематик (наукових спеціальностей), n - кількість наукових спеціальностей, $similarity$ – метрика схожості між науковими спеціальностями. У якості метрики було обрано коефіцієнт Жаккара, що також використовувався у нашій попередній роботі [7]:

$$similarity(S1, S2) = \frac{C}{k1+k2-C} \quad (2)$$

де $S1$ – 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.

Множини відсортовані за спаданням ступенів належності. За рахунок перерахунку ступенів належності може змінитись порядок тематик і це може вплинути на наступні етапи відкидання непопулярних тематик.

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. G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” IEEE transactions on knowledge and data engineering, vol. 17, no. 6, pp. 734–749, 2005.

4. Kuncheva L. Combining pattern classifiers: methods and algorithms / L. Kuncheva. – John Wiley & Sons, 2004. – 350 p.

5. Sun, C., King, T. J., Henville, P., Marchant R. Hierarchical Word Mover Distance for Collaboration Recommender System / C. Sun, T. J. King, P. Henville, R. Marchant // Australasian Conference on Data Mining. Communications in Computer and Information Science, Springer - 2018. – V. 996. – P.289-302.

6. Петричко М. В. Автоматична категоризація науковців за тематикою досліджень на основі профілей в Google Scholar [Електронний ресурс] / М. В. Петричко, С. Д. Штовба // ВНТУ. – 2018. – Режим доступу до ресурсу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2018/paper/view/5427/4433.

7. Петричко М. В. Статистичний підхід до оцінювання подібності наукових спеціальностей в системі Dimensions [Електронний ресурс] / М. В. Петричко, С. Д. Штовба // ВНТУ. – 2019. – Режим доступу до ресурсу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2019/paper/view/7821/6411.

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: 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: roman.tylets@gmail.com.*

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 : 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.*