УДК 681.3

С. Д. Штовба д.т.н., проф., М.В. Петричко, аспірант

АВТОМАТИЧНА КАТЕГОРИЗАЦІЯ НАУКОВЦІВ ЗА ЇХ ІНТЕРЕСАМИ В GOOGLE SCHOLAR

З розвитком науки і техніки наукові спільноти все більше і більше об'єднуються в різноманітні онлайн 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 результати пошуку формуються за буквальним співпадінням. Наприклад, видачі для "fuzzy set" та "fuzzy sets" будуть різними, не говорячи вже синонімічні інтереси типу "fuzzy evidence" та "fuzzy inference". Також не враховує Google Scholar і сукупність інтересів у профілі науковця. Таким чином, пошукові та аналітичні сервіси за велетенським масивом профілів науковців в Google Scholar є досить примітивними.

Подібним питанням займається бібліометрика української науки – єдине джерело українських науковців, що класифіковані за галуззями науки. Проте в існуючій системі дана проблема вирішується суб'єктивно, тобто, задача класифікації науковця виконується людиною. Окрім цього відношення до галузі науки одиничне хоча науковець може бути обізнаним у декількох галузях науки. Тому актуальною є задача побудови профілю науковця таким чином, щоб його можна було б легко знайти або порівняти з іншим. Найпростіший спосіб побудови профілю – знаходження розподілу наукових інтересів по деякій специфікації наук.

Існують задачі, що можуть бути вирішені при наявності бази категоризованих науковців. Одна з таких задач – формування пулу рецензентів для оцінювання дисертації, статті чи запиту на фінансування наукового проєкту. Роботи над цією тематикою мали місце у [1-2].

Постановка задачі. Вважатимемо відомими: $\boldsymbol{W = (w_1, w_2, \dots, w_n)}$ – перелік ключових слів та словосполучень, якими описано наукові інтереси науковця в його профілі в Google Scholar; $\boldsymbol{T = (t_1, t_2, \dots, t_m)}$ – перелік можливих класів, тем – наукових спеціальностей за деякою класифікацією наук; $\boldsymbol{D = (d_1, d_2, \dots, d_k)}$– множина розмічених текстів – множина публікацій кожна з яких віднесена до однієї або декількох тем (наук) $\boldsymbol{T}$; $\boldsymbol{R(D, T) \subset D \times T}$ – відношення, яке описує належність публікацій до наук; $\boldsymbol{R(d_j, t_p) = 1}$, якщо публікацію $\boldsymbol{d_j}$ віднесено до класу $\boldsymbol{t_p}$, $\boldsymbol{j = \overline{1, k}}$, $\boldsymbol{p = \overline{1, m}}$. Задача полягає у знаходженні тем (наук) з $\boldsymbol{T}$, яким відповідає множина ключових слів (інтересів науковця) $\boldsymbol{W}$. При цьому вказується не лише сам факт належності, але і ступінь належності. Таким чином на виході отримуємо нечітку множину $\boldsymbol{\widetilde{W}}$ на підмножині універсальної множини $\boldsymbol{T}$.

Пропонується наступний шлях розв'язання поставленої проблеми. Категоризацію здійснюватимемо навколо деякої специфікації наук. Для кожного науковця, використовуючи його інтереси, знайдемо розподіл інтересів по наукам деякої специфікації. Спрощений опис алгоритму: розрахувати обсяги наукових спеціальностей $\boldsymbol{N_1, N_2, \dots, N_m}$; знайти $\boldsymbol{Q}$ – кількість документів з $\boldsymbol{D}$, в яких є науковий інтерес $\boldsymbol{w_i}$; знайти кількість $\boldsymbol{t_1(w_i), t_2(w_i), \dots, t_m(w_i)}$ документів наукових спеціальностей з науковим інтересом $\boldsymbol{w_i}$; розраховуємо частоти входжень наукового інтересу $\boldsymbol{w_i}$ у наукові спеціальності $\boldsymbol{\gamma_p(w_i) = \frac{t_p(w_i)}{Q_p}}$; якщо наукових інтересів більше ніж один, усереднюємо ступені належності по кожному інтересу і отримуємо розподіл інтересів науковця. Окрім вище описаних кроків виконується ще фільтрування шумів та викидів. Категоризація науковців реалізована під Australian and New Zealand Standard Research Classification з використанням інформаційних ресурсів системи Dimensions.

Висновки. Запропонована модель дозволяє вирішувати проблему категоризації науковця за спеціальностями та галузями знань на основі переліку їх інтересів, які обрано у вільному стилі без прив'язки до будь-якої системи класифікації наук.

Література

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.