УДК 009.04

Р. Ш. Гасанова, асп.

АВТОМАТИЗАЦІЯ ОЦІНКИ ДИСЕРТАЦІЙНИХ РОБІТ

Відпрацьовано метод оцінки дисертаційних робіт з використанням нових наукометричних показників. Для виміру тематичної близькості розділів підготовлено модель, яка отримана за допомогою метрики косинуса. Одночасно запропоновано метод, що дозволяє автоматично визначити актуальність тем дисертацій. Тут розглядається близькість між стислою анотацією вступу й посилань, що відповідають вступу. Далі пропонується оцінка актуальності за розподілом літературних посилань по роках.

Вступ

Розглянемо історію появи наукометрії як науки. Останнім часом збільшення кількості наукових праць почало турбувати наукові інститути. У центрі дослідження наукометрії, створеної в 60-х роках XX століття, стоять: кількість дослідників, видання, посилання, витрати на наукові дослідження тощо. У розвиток цієї науки внесли свій внесок такі вчені, як В. В. Налімов, З. М. Мулченко, І. В. Маршакова. У наш час у Філадельфії (США) існує інститут, заснований 1961 року «батьком» наукометрії Юджином Гар-філдом, під назвою Корпорація Томсона (Інститут наукової інформації). У наукометрії для оцінки наукових праць учених були використані різні наукометричні показники (імпакт-фактор, індекс Хирша й т. ін). До сьогоднішнього дня для об'єктивного оцінювання дисертаційних робіт не використовувалися ніякі наукометричні показники. У цій роботі пропонуються моделі саме для оцінки дисертаційних робіт у різних аспектах. Ясно, що для оцінки однієї дисертації з боку ВАКу часом недостатньо декількох місяців. І ніхто не зможе запевнити нас у збереженні об'єктивності. Для відносного полегшення цього процесу, а також для багаторазового зменшення витраченого на це часу запропоновано автоматизувати процес оцінювання дисертації.

Основна частина

Відомо, що оцінювання дисертаційних робіт потребує від учених багато зусиль і часу. Як зазначено вище, об'єктивність є однією з найважливіших умов в оцінюванні. Для суттєвого спрощення цього процесу потрібно описати формальну структуру дисертаційних робіт.

Позначимо через D елементи множини дисертації й подамо її формальну структуру такий чином:

$$D = \{I, C, R, L\},$$

де, I — вступ, C — множина розділів, R — висновки, L — список літератури.

Рис. 1. Графічне подання структури дисертації

Припустимо, що задано сукупність розділів $C = \{c_1, c_2, \ldots, c_n\}$, як показано на рис. 1. $T = (t_1, t_2, \ldots, t_m)$ — терміни, що зустрічаються в сукупності $C = \{c_1, c_2, \ldots, c_n\}$.

Перше наше завдання полягає у визначенні близькості між розділами. Для текстових даних однією з найпоширеніших мір близькості є метрика косину-

са [1]. Із цією метою в рамках моделі векторного простору (Vector Space Model) кожному терміну t_t зіставляється деяка позитивна зважена функція w_{it} [2]. Таким чином, кожний розділ буде подано у вигляді m-вимірного вектора $c_i = (w_{i1}, w_{i2}, \ldots, w_{im})$, $i = \overline{1,n}$. Вага w_{it} терміна t_t залежить від частоти його появи в конкретному розділі, що визначається формулою TF × IDF (Term Frequency × Inverse Document Frequency):

$$w_{it} = tf_{it} \log\left(\frac{n}{n_t}\right), \tag{1}$$

де, tf_{it} — частота появи терміна t_t в розділі c_i, n — загальна кількість розділів у дисертації, n_t — кількість розділів, у яких присутній термін t_t.

Для визначення змістової близькості між розділами в дисертаційній роботі доцільно вибрати метрику косинуса [3]. Відповідно до цього, метрика міри подоби розділів c_i і c_j визначається з формули

$$\cos(c_i, c_j) = \frac{\sum_{t=1}^{m} w_{it} w_{jt}}{\sqrt{\sum_{t=1}^{m} w_{it}^2} \sqrt{\sum_{t=1}^{m} w_{jt}^2}}, \quad i, j = \overline{1,n}. \tag{2}$$

Залежно від значення $\cos(c_i, c_j)$ можна говорити наскільки розділи близькі між собою за тематикою.

На наступному етапі спробуємо оцінити актуальність тем дисертаційних робіт. Для цього введемо такі позначення: A_I — стисла анотація вступу, L_I — список посилань, що відповідають вступу (рис. 2).

Рис. 2. Оцінка актуальності теми дисертаційної роботи

Нехай $T = (t_1, t_2, \ldots, t_m)$ буде набором термінів, що зустрічаються в A_I і L_I.

Визначимо близькість між множинами A_I й L_I. Для текстових даних однією з найпоширеніших мір близькості є метрика косинуса [1]. Із цією метою в рамках моделі векторного простору (Vector Space Model) кожному терміну t_k зіставляється деяка зважена функція w_k [2]. Таким чином, множини A_I й L_I будуть подані у вигляді m-вимірних векторів $A_I = (w_1^A, w_2^A, \ldots, w_m^A)$ і $L_I = (w_1^L, w_2^L, \ldots, w_m^L)$ відповідно.

$$w_k^{A,L} = tf_k^{A,L} \log\left(\frac{n}{n_k}\right),$$

де, $tf_k^{A,L}$ — кількість появ терміна t_k в множинах A_I і L_I відповідно; n — загальна кількість множин (у нашому випадку 2); n_k — кількість множин, у яких зустрічається термін t_k.

Близькість між A_I і L_I визначається за такою формулою:

$$\cos(A_I, L_I) = \frac{\sum_{k=1}^{m} w_k^{A_I} w_k^{L_I}}{\sqrt{\sum_{k=1}^{m} \left(w_k^{A_I}\right)^2} \sqrt{\sum_{k=1}^{m} \left(w_k^{L_I}\right)^2}}.$$

Далі в списку літератури здійснюється поділ по роках і множина L_I розділяється на 3 підмножини: $L_I = \left(L_I^5, L_I^{6-10}, L_I^{11-}\right)$.

Тут, L_I^5 — список літератури за останні 5 років, L_I^{6-10} — список літератури за останні 10 років з відніманням 5 останніх років, L_I^{11-} — список літератури, що передує останнім десяти рокам.

На наступному етапі визначається близькість між цими підмножинами й множиною A_I, тобто обчислюються $\cos\left(A_I, L_I^5\right)$, $\cos\left(A_I, L_I^{6-10}\right)$ і $\cos\left(A_I, L_I^{11-}\right)$. Близькість із найбільшим результатом називається актуальною, відносно актуальною й неактуальною відповідно.

Висновки

1. Збіг розділів у дисертаційній роботі можна визначити за такою моделлю:

$$\cos\left(c_i, c_j\right) = \frac{\sum_{t=1}^{m} w_{it} w_{jt}}{\sqrt{\sum_{t=1}^{m} w_{it}^2} \sqrt{\sum_{t=1}^{m} w_{jt}^2}}, \quad i, j = \overline{1, n}.$$

2. Актуальність теми дисертаційної роботи можна визначити за такою моделлю:

$$\cos\left(A_I, L_I\right) = \frac{\sum_{k=1}^{m} w_k^{A_I} w_k^{L_I}}{\sqrt{\sum_{k=1}^{m} \left(w_k^{A_I}\right)^2} \sqrt{\sum_{k=1}^{m} \left(w_k^{L_I}\right)^2}}.$$

СПИСОК ЛІТЕРАТУРИ

1. Алгулиев Р. М. Аннотирование текстовых документов с определением скрытых тематических разделов и информативных предложений / Р. М. Алгулиев, Р. М. Алыгулиев // Автоматика и вычислительная техника. — 2007.

2. Salton G. A vector space model for automatic indexing / G. Salton, A. Wong, and C. S. Yang // Communication of the ACM. — 1975.— Vol. 18, No 11. — P. 613—620.

3. Alguliev R. M. Effective summarization method of text documents / R. M. Alguliev, R. M. Aliguliev // Proc. of the 2005 IEEE/WIC/ACM : International Conf. on Web Intelligence (WI'05). — France. : Compiegne University of Technology, 2005, September 19—22. — P. 264—271.

Рекомендована кафедрою інтелектуальних систем

Надійшла до редакції 8.09.08
Рекомендована до друку 20.10.08

Гасанова Рахіль Шабан кизи — аспірантка, науковий співробітник Інституту інформаційних технологій.

Національна Академія наук Азербайджану, м. Баку