УДК 51.001.57+004.652.4+004.827

Р.Ю. НОГА[*]

МЕТОД ФОРМУВАННЯ НАУКОВИХ ШКІЛ НА ОСНОВІ АНАЛІЗУ ЕЛЕМЕНТІВ ПУБЛІКАЦІЙ

[*]Національний університет «Львівська політехніка», Львів, Україна

Анотація. *У статті розроблено метод визначення елементів наукової публікації та об'єднання авторів публікацій у наукові школи.*
Ключові слова: *наукова публікація, наукові школи, об'єднання авторів.*

Аннотация. *В статье разработан метод определения элементов научной публикации и объединения авторов публикаций в научные школы.*
Ключевые слова: *научная публикация, научные школы, объединения авторов.*

Abstract. *Method for determination of scientific publications elements and associations of authors into scientific schools was developed in this paper.*
Keywords: *scientific publication, scientific schools, associations of authors.*

1. Вступ

Переробка інформації, представлена у вигляді текстів природною мовою, має багато аспектів. Сюди відносяться такі види інформаційних процесів, як розуміння текстів, їх переклад, стиснення семантичної інформації. Особливе значення має останній тип переробки; сюди відносяться класифікація та індексування документів, їх анотування та реферування.

Останнім часом серед науковців, редакторів наукових журналів тощо постає проблема кластеризувати публікації за науковими школами з метою визначення фаховості статті, споріднених публікацій та ін. Проте поняття «наукова школа» є неформалізованим.

Тому метою статті є розроблення методу формування наукових шкіл на основі аналізу публікацій.

2. Аналіз літературних джерел

Оскільки основою формування наукової школи є аналіз текстів, розглянемо методи видобування інформації з тексту.

Процес реферування текстової інформації на сьогоднішній день є дуже актуальним, не дивлячись на величезну кількість робіт. У першу чергу, це викликано постійним зростанням неструктурованих даних Веб-ресурсів, підвищенням вимог до продуктивності та часу відклику на запит. Крім того, реферування є невід'ємною частиною сучасного видавничого процесу. Будь-яке видання, чи це монографія, підручник, аналітичний огляд тощо, завжди випереджується вторинним документом (рефератом або анотацією). Реферування використовується не тільки для економії часу при ознайомленні з великою кількістю джерел, але й з метою пришвидшення повнотекстового пошуку по множині документів, оскільки обсяг реферату у декілька разів менший, ніж обсяг вхідного документа чи їх множини [1].

Яким чином можна автоматизувати процедуру стискання семантичної інформації для отримання реферату? Мета процедури автоматизованого реферування – виділити з тексту документа найважливіші положення, які найповніше розкривають суть цього тексту.

Серед таких положень для наукових публікацій можна визначити такі, як автор видання, наукова установа, тема, ключові слова. Саме визначення цих чотирьох елементів

дає змогу зробити швидкій пошук контенту, інтегрування текстової та структурованої інформації.

На сьогоднішній день методи автоматичного аналізу текстів (text mining) широко використовуються в різних галузях науки. Використовують три основні підходи: підхід, заснований на аналізі назв об'єктів, які зустрічаються в текстових документах і так званий повний та поверхневий парсинг.

Повний парсинг базується на описі мови за допомогою формальних граматик. Основним недоліком такого методу є високі вимоги до часу виконання. У зв'язку з цим цей метод має обмежену область застосування. Як приклад систем аналізу текстів, які працюють за принципом повного парсингу, можна навести PathwayStudio [5] і GeneScene [6].

Поверхневий парсинг оснований на витягуванні формалізованої інформації з тексту з використанням часткових зв'язків між словами за допомогою набору спеціальних шаблонів та правил. На цьому методі основані такі системи, як SUISEKI [7], Chilibot [8] та ін.

Однак існуючі системи аналізу текстів орієнтовані на певні предметні області [1–4] і тому не можуть бути використані для аналізу наукових публікацій певної наукової установи.

3. Метод формування наукових шкіл. Виділення складових елементів наукової публікації

Введемо поняття наукової школи.

Науковий напрям – це сфера наукових досліджень наукового колективу, спрямованих на вирішення певних значних фундаментальних проблем.

Наукова школа – науковий колектив, діяльність якого спрямована на вирішення проблем наукового напряму.

У цьому дослідженні наукова школа *Sch* визначатиметься множиною наукових публікацій *P*, які характеризуються множиною ключових слів *Key*, множиною авторів *Author* та множиною основоположників школи *Main*:

$$Sch = \langle Key, Author, Main \rangle, Main \in Author .$$

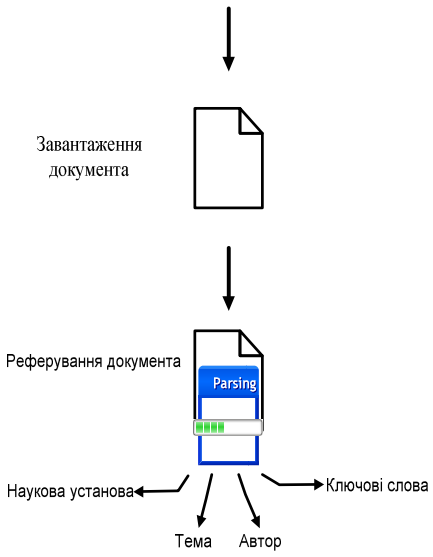

Рис. 1. Виділення інформації з контенту

Наукові публікації *P* подаються у вигляді текстових даних ***Nd*** та Веб-сайтів ***Wb***.

Необхідні кроки для виділення з контенту необхідної нам інформації для подальшої роботи з нею подані на рис. 1.

Виділення класифікаційних ознак наукової публікації здійснюватиметься на основі семантичної мережі.

Семантична мережа – це структура для подання знань у вигляді вузлів, з'єднаних дугами. Семантична мережа, побудована на основі аналізу термів напівструктурованого джерела інформації Γ, подається як двійка:

$$\Gamma = \{V, D\},$$

де $V = \{v_i\}$ – множина вершин (вузлів мережі), $D = \{d_j\}$ – множина дуг. Дуги між елементами

визначають взаємозв'язки між вершинами і задають послідовність пошуку концептів (їх важливість).

Побудуємо функцію трансформації напівструктурованого тексту та Веб-сайтів у вигляді семантичної мережі:

$S(E) \to N, E \in \mathbf{Wb} \vee E \in \mathbf{Nd}$ – для Веб-сайтів, текстових даних.

Результатом операції S є неорієнтований граф.

Між двома будь-якими елементами Y_i, Y_j словника даних Dic, $Y_i \in Dic, Y_j \in Dic$ існує відображення

$$\forall Y_i : \exists n, \Gamma^n(Y_i) = \left\{ Y_i, i = \overline{1, M} \right\},$$

де $\Gamma(Y_i) = \left\{ Y_j : \exists S(Y_i, Y_j) \vee S(Y_j, Y_i) \right\}$.

Формуються підграфи для кожного Y_i, такі, що в підпункті вузол вихідного параметра один, а інші вузли – це вхідні поняття, що описують обмеження на атрибути $\left\{ X_k, 1 \le k \le N \right\} \leftarrow Y_i \leftarrow \left\{ X_i, 1 \le l \le N \right\}$, тут $X_k \leftarrow Y_i = S(X_k, Y_i) : Y_i \leftarrow X_k = S(Y_i, X_k)$. Крім цього, у граф так само входять усі вхідні поняття, які використовуються як обмеження: $\forall Y_i : \Gamma'^k(Y_i)$, де

$$\Gamma'(Y_i) = \left\{ X_j : \exists S(Y_i, X_j) \vee S(X_j, Y_i) \right\}.$$

$$\Gamma'^2(Y_i) = \Gamma'(Y_i, \Gamma'(X_j)) = \left\{ X_k : \exists S(Y_i, X_k) \vee S(X_k, X_i) \right\}.$$

Дуги між вузлами $\Gamma'^k(Y_i)$ визначаються на основі існуючих відношень між поняттями S^k і підграфи даного типу можна визначити як $G'(Y_i) = \left\langle \Gamma'^k(Y_i), S^k \right\rangle$.

Друга множина підграфів визначається як вузли з вихідних понять, і відношення між ними $G''(Y_i) = \left\langle \Gamma''^n(Y_i), S^n \right\rangle$, де $\Gamma''(Y_i) = \left\{ Y_j : \exists S(Y_i, Y_j) \right\}$.

Для всіх підграфів $G'(Y_i)$ формується запит, що забезпечує всю вибірку примірників Y_i.

Для підграфа $G''(Y_i)$ формується запит, забезпечує вибірку примірників Y_i на основі даних по Y_j, отриманих на попередньому кроці.

Наступні функції виконуються в автоматичному режимі:

– визначення тематичних рубрик документа;

– визначення об'єктів на основі онтологічного описання;

– формування пошукового образу документа;

– формування частотного словника ключових слів і словосполучень.

Результатом побудови семантичної мережі є розроблення тезауруса.

Тезаурус – це $Th = < T, R >$, де T – множина термінів, а R – множина відношень між цими термінами. Множини T і R скінченні. Термін – це слово або словесний комплекс, який співвідноситься з поняттям певної організованої області знань (науки, техніки), що вступає в системні відношення з іншими словами і словесними комплексами й утворює разом з ними в будь-якому окремому випадку та у певний час замкнену систему, яка відрізняється високою інформативністю, однозначністю, точністю й експресивною нейтральністю.

Тезаурус – структура лінійно пов'язаного подання слів і їхніх значень, призначена для співставлення концептуальних визначень у контексті слова [1]. Множина термінів тезауруса відповідає множині концептів онтології О.

Приклад тезауруса області наукових досліджень поданий на рис. 2.

Структура тезауруса визначена стандартами ANSI Z39.19, ISO 2788-1986, ISO 5964-1985, ГОСТ 7.25-2001, ГОСТ 7.24-90. Для врахування ефектів, пов'язаних з розбіжністю

суб`єктивних знань приймача і передавача в комунікаційних процесах, що є наслідками різних обсягів знань у ПО, використовують тезаурусну модель, яка зв'язує семантичні властивості інформації зі здатністю користувача сприймати інформацію.

людина (ім'я (STRING), по батькові (STRING), прізвище(STRING), рік народження (INTEGER)):
співробітник (…, посада (STRING), працює в (підрозділ), ідентифікаційний код (INTEGER)):
науковий співробітник (…, науковий ступінь (STRING), працює за темою (тема), науковий стаж (DATE), публікації (публікація));
інженер(…, має кваліфікацію (STRING));
аспірант (…, рік вступу (DATE), науковий керівник (науковий співробітник), публікації (публікація));
підрозділ (назва (STRING), керівник (співробітник)):
інститут (…, адреса (STRING));
відділ (…, належить до (інститут));
лабораторія (…, належить до (відділ));
тема (шифр (STRING), назва (STRING), керівник (науковий співробітник), дата початку (DATE), дата закінчення (DATE), виконавці (співробітник)):
комплексна тема(…, складається з (тема)) ;
публікація (назва (STRING), автори (людина), рік публікації (STRING), мова (STRING), кількість сторінок (INTEGER)):
наукова стаття (…, УДК (STRING), анотація (STRING), назва видання (STRING));
монографія (…, рецензент (науковий співробітник), назва видавництва (STRING));
тези конференції (назва конференції (STRING), дата ПО ведення (DATE), місце ПО ведення (STRING)).

Рис. 2. Тезаурус онтології наукових досліджень

Алгоритм формування бази даних характеристик публікації передбачає такі кроки:

Крок 1. Наукова стаття, подана як структурована текстова інформація, розбивається на речення та слова.

Крок 2. Відкидаються слова, що містять менше трьох символів.

Крок 3. Здійснюється класифікація слів шляхом видалення з загального списку слів, які містяться в базі даних «Стоп-слова» та неінформативних слів і словосполучень.

Крок 4. Формується загальний список слів у документі, при цьому зберігається інформація про їх форматування та місце в тексті.

Крок 5. Загальний список слів модифікується у процесі стеммінгу, тобто відкидаючи закінчення слів, ми також видаляємо однакові слова з бази даних, але збільшуємо значення, що відповідає за кількість вживань цього слова в тексті, а ваги, що були попередньо присвоєні цим словам, додаються. Таким чином, утворюється база даних «Ключові слова тексту».

Крок 6. Автори статті та їх наукові установи шукаються на початку файлу за ознакою форматування.

4. Кластеризація наукових публікацій

Нехай ми маємо деяку публікацію P. Після побудови семантичної мережі даної публікації ми отримуємо такі елементи:
Автор =>A;
Наукова установа =>B;
Тема =>C;

Ключові слова =>D.

Після того, як ми провели аналіз даних та отримали необхідну інформацію, можемо приступити до кластеризації публікації.

Кластеризація – це автоматичне розбиття елементів деякої множини на групи. Кластеризацію проводитимемо методом k-найближчих сусідів.

Метод найближчих сусідів полягає у виконанні таких кроків.

1. Задаємо кількість сусідів k.

Оскільки ознаки кластеризації (автор, наукова установа, тема, ключові слова) невпорядковані, то використовуватимемо метрику d ізольованих точок:

$$l(X.x, Y.x) = \begin{cases} 1, X.x = Y.x \\ 0, X.x = Y.x \end{cases},$$

$$d(X, X_i) = \sum_i^p l(X.A_i, Y.A_i) + \sum_j^r l(X.D_j, Y.D_j) + \sum_t^w l(X.B_t, Y.B_t) + l(X.C, Y.C),$$

де p – кількість авторів обох статей, r – сумарна кількість ключових слів, w – сумарна кількість наукових установ, $X.A_i$ – значення автора з номером i для наукової статті X і т.д.

2. Для кожного об'єкта знаходимо його k найближчих сусідів. Об'єкт X_i називається найближчим сусідом об'єкта X, якщо $d(X_i, X) = \max_i d(X_i, X), i = \overline{1, N}$, де N – кількість публікацій.

3. Об'єкт X зараховується до того класу, до якого належить більшість з його k сусідів.

Якщо об'єкт не зарахований до жодного з кластерів, то шукаються слабкі зв'язки об'єкта з кластером.

Слабким назвемо зв'язок між об'єктами X_i та X, якщо значення відстані між ними менше, ніж третина від максимальної:

$$d_s(X, X_i) \le \frac{\max d(X, X_i)}{3}.$$

Продемонструємо, яким чином здійснюється формування наукових шкіл.

Нехай маємо деякі публікації P1 та P2.

Спочатку виділяємо інформацію про автора, наукову установу, ключові слова та тему.

Ми отримаємо множини P1 та P2 з деякими характеристиками:

$$P1 = \begin{cases} A = a11,\ a12 \\ B = b1 \\ C = c1 \\ D = d11,\ d12,\ d13 \end{cases} \quad ma \quad P2 = \begin{cases} A = a21,\ a22 \\ B = b2 \\ C = c2 \\ D = d21,\ d22,\ d23 \end{cases},$$ де $a11, a12$ – автори і т.д.

Тепер нехай маємо публікації P3 та P4. Робимо аналогічне витягування інформації.

Отримаємо таке:

$$P3=\begin{cases} A=a31,\ a11 \\ B=b31,b1 \\ C=c3 \\ D=d31,\ d32,\ d33,\ d13 \end{cases} \quad ma \quad P4=\begin{cases} A=a41,\ a22 \\ B=b41,b2 \\ C=c4 \\ D=d41,\ d42,\ d43,\ d22 \end{cases}.$$

Визначаємо кількість спільних елементів для кожної з публікацій.

Публікації P3 та P4 мають деякі спільні характеристики з P1 та P2, а саме: це $a11$ (автор), $b1$ (наукова установа), та $d13$ (ключові слова). Так само в P4.

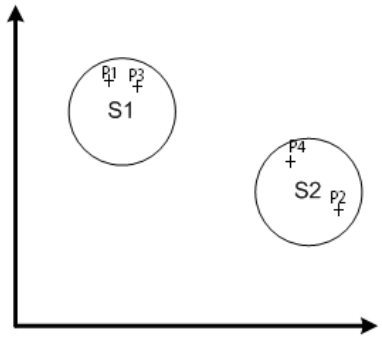

Рис. 3. Представлення шкіл зі статтями

Ми маємо чотири множини, розбиті за характеристиками. Тепер можемо об'єднати множини P1..P4 за спільними характеристиками. Так як P1 та P3, а також P2 та P4 мають спільних авторів, наукові установи, де вийшли публікації, та ключові слова, ми отримаємо кластери $\{P1, P3\}$ та $\{P2, P4\}$:

$$P1,P3=\begin{cases} A=a11 \\ B=b1 \\ D=d13 \end{cases} \quad ma \quad P2,P4=\begin{cases} A=a22 \\ B=b2 \\ D=d22 \end{cases}.$$

Отримані групи і будуть формувати школи *Sch*. Отже,

$$Sch1=\{P1, P3\} \text{ та } Sch2=\{P2, P4\}.$$

Тепер уявімо собі, що в нас є деяка публікація P5. Нехай після виділення елементів публікації ми отримаємо таку множину ознак:

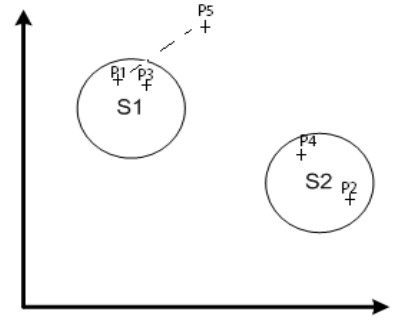

Рис. 4. Представлення шкіл та слабкий зв'язок зі статтею

$$P3=\begin{cases} A=a51,a51 \\ B=b51,b52 \\ C=c5 \\ D=d51,\ d52,\ d53,\ d13 \end{cases}.$$

Ми бачимо, що у множини P5 у нас є спільне з P1 лише одне ключове слово. Ми відносимо P5 до школи S1. Зв'язок P5 та S1 є «слабким», відносити P5 в школу S1 не будемо, тільки зв'яжемо.

Слабкий зв'язок необхідно залишити з тих міркувань, що у майбутньому не виключено, що P5 буде мати спільні характеристики з іншими публікаціями і створиться власна школа S3.

Для випадків, коли ми маємо слабкі зв'язки, можна застосувати метод визначення спільних ознак у назві публікації.

5. Метод визначення спільних ознак у назві публікації

Нехай маємо деякі назви C1,C2,C3. Для прикладу:

C1=«Пошук та збереження інформації за допомогою пошукової системи».

C2=«Перегляд та збереження файлів у файловій системі».

C3=«Пошук інформації у всесвітній мережі інтернет».

Умовно розіб'ємо назви на дві частини: праву та ліву. Розбиття здійснюватиметься шляхом симетричного поділу по довжині. Вважатимемо, що ліва частина є більш інформативно важливою, ніж права.

Розіб'ємо теми на ліву та праву частини й виберемо спільне. При цьому слід не брати до уваги слова-коннектори, такі як «і, та» і т.д. При цьому не слід відкидати слова, написані великими літерами: це може бути абревіатура. Також здійснюється відсікання закінчень.

Тоді отримаємо:

С1л=С3л=«пошук, інформація».

С1л=С2л=«збереження».

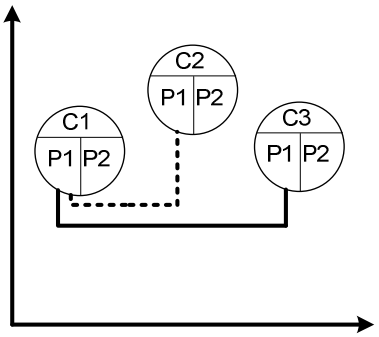

Рис. 5. Представлення спільних ознак у назві публікації

Зв’язок, який утворюється між публікаціями, для яких співпадає більше половини слів у лівій частині назви, назвемо сильним зв’язком назв.

Отже, оскільки С1л та С3л мають два спільних слова, то між публікаціями Р1 та Р3 утворюється сильний зв’язок назв.

Відповідно в назвах С1л та С2л утворено слабкий зв’язок назв.

Такі зв'язки між темами можна використовувати для додаткового навантаження зв'язків між публікаціями, що, у свою чергу, може вплинути на прийняття рішення, в яку із існуючих шкіл відносити публікацію, чи залишати її для створення нової школи.

5. Висновки

У статті запропоновано метод визначення ознак наукових публікацій та їх кластеризації. Кластеризація використовується для формування інформації про наукові школи. Розроблено метод визначення зв’язку між публікацією та школою.

За допомогою такого підходу ми можемо відстежувати, які школи стрімко розвиваються і які занепадають, за якими характеристиками поповнюється школа, та проаналізувати перспективні теми і проблеми.

Також за допомогою шкіл, сформованих подібним шляхом, значно оптимізовується пошук потрібної інформації. Так, для прикладу, якщо користувач шукатиме якусь інформацію, нехай $a11$, тоді система видасть усю спільну інформацію з $a11$, тобто школу S1.

СПИСОК ЛІТЕРАТУРИ

1. Salton G. Automatic Text Structuring and Summarization / G. Salton // Information Processing & Management. – 1997. – Vol. 33, N 2. – P. 193 – 207.

2. Mani I. The Tipster Summac Text Summarization Evaluation / I. Mani // Proc. 9th Conf. European Chapter of the November 2000. – 2000. – P. 118 – 121.

3. Mani I. Summarizing Similarities and Differences Among Related Documents / I. Mani, E. Bloedorn // Information Retrieval. – 1999. – Vol. 1, N 1. – P. 35 – 67.

4. Radev D.R. Generating Natural Language Summaries from Multiple Online Sources / D.R. Radev, K.R. McKeown // Computational Linguistics. – 1998. – Vol. 24, N 3. – P. 469 – 500.

5. Carbonell J.G. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries / J.G. Carbonell, J. Goldstein // Proc. 21st Int’l ACM SIGIR Conf. Research and Development in Information Retrieval. – New York: ACM Press, 1998. – P. 335 – 336.

6. Ando R.K. Multidocument Summarization by Visualizing Topical Content / R.K. Ando // Proc. ANLP/NAACL 2000 Workshop on Automatic Summarization. – 2000. – P. 79 – 88.

Стаття надійшла до редакції 11.12.2012