УДК 004.652.4+004.827

Н. Б. Шаховська, Р. Ю. Нога

РОЗРОБЛЕННЯ АРХІТЕКТУРИ СИСТЕМИ АНАЛІЗУ ФУНКЦІОНУВАННЯ НАУКОВИХ ШКІЛ

In this paper the method of creating of science school based on scientific publications. The system architecture of saving and analysis of publications is given.

Keywords: *system architecture, clustering, scientific school.*

Розроблено методи формування наукової школи на основі наукової публікації, архітектуру системи збереження та аналізу публікацій.

Ключові слова: *архітектура системи, кластеризація, наукова школа.*

Від розвитку та цілеспрямованої наукової роботи значною мірою залежить науковий імідж університету і якість підготовки студентів. Тлумачний словник [5] визначає наукову школу як напрям у науці, пов'язаний єдністю спільних поглядів, наступністю принципів і методів. За визначенням К. Ланге, наукова школа – це неформальний науковий колектив, сформований навколо відомого вченого на базі наукової установи, який поєднує з метою колективної розробки певної наукової ідеї, проблеми, напряму окремі наукові колективи.

Наукова школа – це співдружність людей, що сформувалася під егідою особистості – вченого-лідера, який має ідеї, теми для розробки. Одним з найвагоміших результатів функціонування наукової школи є наукові публікації її учасників.

Стаття присвячена розробленню засобів кластеризації наукових статей з метою формування ознак наукової школи і прогнозування динаміки її розвитку.

Актуальність роботи. Передусім визначимо, за якими показниками доцільно здійснювати аналіз публікацій та визначати їх приналежність до тої чи іншої наукової школи.

Серед чинників ефективного функціонування наукових шкіл виділяють такі [6]:

- визначення наукового напряму, актуальної профільної наукової теми, перспективи її розвитку;
- формування наукових підрозділів (інститут, відділ, лабораторія, центр) при університеті, факультетах, кафедрах;
- формування наукових колективів, ретельне планування наукових досліджень;
- створення сучасної матеріально-технічної дослідницької бази;
- наявність докторантури, аспірантури, інституту здобувацтва;
- опублікування фундаментальних наукових праць: монографій, науково-методичних посібників, статей у фахових виданнях, зокрема міжнародних;
- наявність фахового наукового періодичного видання;
- щорічне проведення наукових заходів: симпозіумів, конференцій, семінарів.

Ознаками наукової школи є наявність наукової спільноти, яка розвивається у часі і просторі; спрямованість на розробку нового, оригінального напряму у науці; спільність наукових інтересів, принципів та методичних підходів у процесі виконання продуктивної програми досліджень; наявність декількох поколінь учених (ланка “учитель–учень”), об’єднаних визнаним лідером; підвищення наукової кваліфікації учасників школи; опублікування наукових результатів (публікації, доповіді).

Критеріями визнання наукової школи [6] є відповідність профільної теми державним пріоритетним напрямам розвитку науки і техніки, програмам Міністерства освіти і науки, Національної асоціації наук, галузі; захист докторських і кандидатських дисертацій за напрямом школи; наявність відкриттів, винаходів; опублікування монографій, публікацій у фахових виданнях, депонування звітів; організація наукових заходів: щорічних міжнародних чи всеукраїнських конференцій, постійно діючих семінарів; створені на базі школи діючі науково-виробничі структури державного рівня.

Отже, ми бачимо, що одним з необхідних критеріїв існування наукової школи є потужна та сучасна база даних різних публікацій та новітніх наукових розробок.

Аналіз літературних джерел та постановка задачі. Серед наявних методів аналізу наукових публікацій є формування онтологій.

Важливу роль у системах аналізу множини текстових документів відіграють Інтернет і World – Wide Web. Використання онтологій для пошуку дає змогу користувачеві сформулювати свій запит на вищому рівні абстракції, ніж це можливо під час пошуку за ключовими словами.

Розглянемо приклади систем, що використовують онтології для роботи з Інтернетом.

OBSERVER (http://siul02.si.ehu.es/~jirgbdat/OBSERVER). Ця система пропонує підхід використання безлічі вже існуючих онтологій для доступу до гетерогенних, розподілених і незалежно розроблювальних репозиторіях даних [1]. Реалізація такого підходу – ідеологія брокера онтологій предметних областей. Передбачається, що є безліч заздалегідь створених онтологій предметних областей, користувачеві необов'язково "підбудовуватися" під конкретну онтологію. Користувач формулює свій запит деякою мовою, у термінах однієї чи декількох онтологій, і брокер "шукає" релевантні інформаційні ресурси, виконуючи транслювання запиту в придатні онтології, а в разі потреби і сполучення декількох онтологій для більш точної відповіді на запит.

OntoSeek [3]. Ця система розроблена для контекстного отримання інформації з он-лайнових "жовтих сторінок" та каталогів продуктів. Система може працювати як з однорідними, так і з неоднорідними каталогами продуктів. Для точної фіксації контексту може бути застосований інтерактивний підхід, коли користувач поступово уточнює зміст ключових слів за допомогою лінгвістичної бази даних WordNet. WordNet – це лінгвістична база даних, що складається із сінсетів (synsets) – груп слів, еквівалентних за змістом. WordNet є водночас і лексичним словником (створеним для декількох європейських мов), і онтологією, що відображає зв'язки між словами у словнику. Опис ресурсу реалізується у вигляді лексичного концептуального графа [1], де вершини відповідають словам, а іменовані дуги – семантичним відносинам між словами (наприклад, відносини типу "частина" або "підклас" та ін.), назви вершин і дуг також беруть із WordNet, під час створення концептуального графа конкретного ресурсу. Знаходження ресурсів, відповідних до запиту користувача, базується на порівнянні онтологій (лексичних концептуальних графів) цих ресурсів. Під час відбору ресурсів, відповідних до запиту користувача, OntoSeek виконує порівняння концептуального графа запиту із існуючими концептуальними графами ресурсів або з частинами цих графів.

OntoSeek має централізований сервер, на якому розміщена база даних лексичних концептуальних графів відомих системі ресурсів, але створює такі графи клієнт.

Підхід, використаний в OntoSeek, відрізняється від підходу, який застосовують у моделі W3C Resourse Description Framework (W3C RDF). У RDF опис структури даних (тобто схема даних у вигляді <subject, predicate, object>), додається у

HTML/XML документ, а не зберігається окремо. Ніяких додаткових умов щодо семантичної узгодженості даних RDF не вимагає.

Онтологічний підхід – практично найкращий для пошуку статей певної наукової школи [2]. Проте онтологій українською мовою є не так багато. Тому використання онтологічного пошуку можливо вже на наступних етапах побудови системи аналізу діяльності наукових шкіл.

Як бачимо, майже за усіма зазначеними напрямками ведуться роботи. Але ці роботи неінтегровані, не передбачають єдиного опрацювання та жорстко прив'язані до моделі даних, що є цілком неприйнятним у контексті просторів даних. Тому проблема формалізації просторів даних є актуальною.

Основний матеріал. Подамо декілька визначень.

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 визначатиметься множиною наукових публікацій Sch, які характеризуються множиною ключових слів Key, множиною авторів Author та множиною основоположників школи Main:

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

Поставлено задачі:

– визначити за публікаціями, які наукові школи функціонують;

– класифікувати нові надходження публікацій за науковими школами;

– прогнозувати динаміку появи нових публікацій наукової школи;

Вхідною інформацією для віднесення публікації до наукової школи є файл з вмістом публікації. З файлу необхідно визначити базові характеристики публікації:

1. Автор(и) публікації (A).
2. Наукова установа (B).
3. Тема публікації (C).
4. Ключові слова (D).
6. Текст статті.

Алгоритм формування бази даних характеристик публікації складається з таких кроків:

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

Крок 2. Відкидаються слова, що містять менше, ніж три символи.

Крок 3. Здійснюється класифікація слів шляхом вилучення зі 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 вилучаємо 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. Так утворюється база даних "Ключові слова тексту".

Крок 6. Авторів статті та їхні наукові установи шукають на початку файлу за ознакою форматування.

Після формування бази даних цієї публікації відбувається розбиття публікацій за науковими школами методом *k*-середніх (*k*-means):

1. Задаємо *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 і для наукової статті X і т.д.

2. Для кожного об'єкта знаходимо його k сусідів. Об'єкт X_i називають 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 зачисляється 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 (наявність хоча б по одному спільному елементу в множинах характеристик публікацій A, B, D):

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

Далі спроектуємо архітектуру системи аналізу функціонування наукових шкіл (рис. 1).

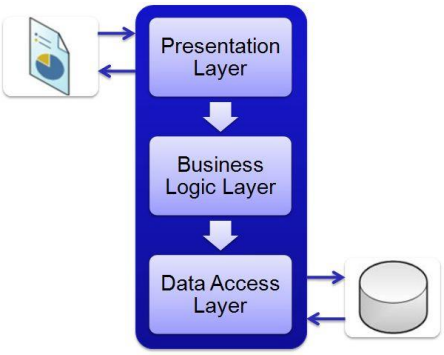


Рис. 1. Архітектура системи.

У цій системі маємо трирівневу архітектуру:

– нижній рівень – робота з даними: опрацювання бази даних наукових публікацій;

– рівень бізнес-логіки, де виконується власне кластеризація наукових публікацій та прогнозування розвитку наукових шкіл;

– рівень презентації даних.

Для збереження елементів публікації спроектуємо базу даних "дата трансфер об'єкти" (рис. 2):

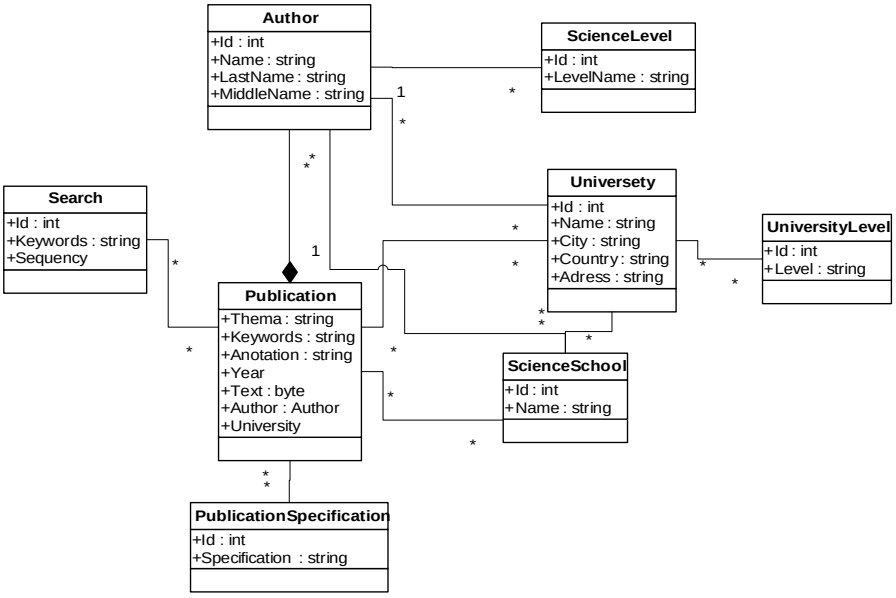


Рис. 2. Дата трансфер об'єкти.

Запис частин публікації, вага яких перевищує мінімально задану, здійснює бізнес-рівень аплікації, зображений на рис. 3.

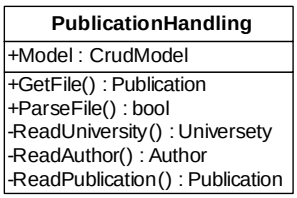


Рис. 3. Бізнес-рівень системи під час витягнення інформації з файлу.

Далі інформація передається на рівень роботи з даними. Оскільки цей рівень відповідає за роботу з базою даних, то там ми і запишемо дані у базу у відповідні таблиці.

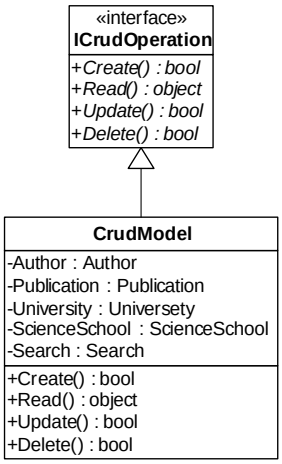
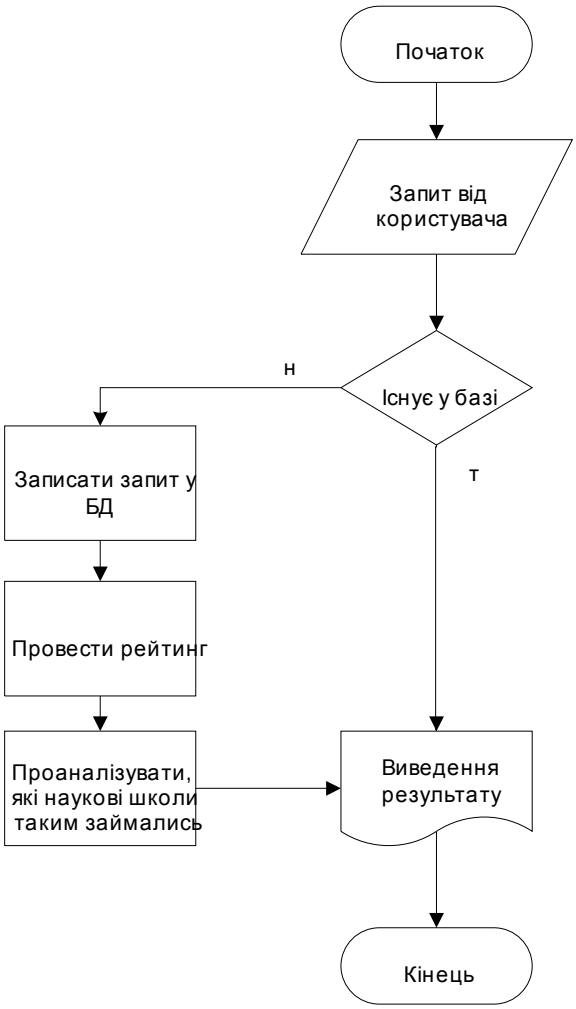


Рис. 4. Рівень роботи з даними.

У базі даних, поданій на рис. 1, є таблиця “Пошук”. Інформація з цієї таблиці є вхідною для прогнозування розвитку наукової школи. Як було зазначено, ще одними важливими критеріями розвитку наукової школи є визначення наукового напряму, актуальної профільної наукової теми, перспективи її розвитку. Маючи запис пошуку користувачами джерел інформації, доцільно розробити аналізатор даних визначення актуальних тематик.

Алгоритм аналізатора має такі кроки. Під час пошуку інформації користувач вводить дані, які він шукає, відповідно здійснюється пошук у базі даних на наявність інформації, яка шукається. За відсутності даних у базі запит користувача записується у таблицю пошук. Таким чином ми будемо мати запис, який свідчить, що публікацію з такою темою ще не розглядали. Також на основі записів у цій таблиці ми можемо зробити рейтинг актуальних пошуків. Далі визначаємо, які наукові школи досліджували раніше подібні теми, та проінформуємо їх про нову актуальну тему. Також користувача, який здійснює пошук, ми можемо проінформувати, які наукові школи займаються подібними темами та/або конкретні люди, до яких можна звернутись. Алгоритм аналізатора поданий на рис. 5.



За роботу цього аналізатора відповідатиме частина бізнес-рівня, що матиме такий вигляд:

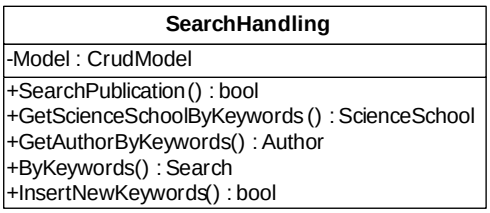


Рис. 5. Бізнес-частина аналізатора для пошуку даних.

ВИСНОВКИ

У статті розроблено метод формування наукових шкіл за результатами аналізу публікацій, а також архітектуру системи аналізу наукових публікацій. Подальші дослідження спрямовані на прогнозування створення наукових шкіл на основі алгоритмів, наведених у попередніх статтях, та дослідження роботи аналізатора як одного з методів прогнозування розвитку наукової школи.

1. *Automatic* Text Structuring and Summarization / Salton G. et al. // Information Processing & Management. – 1997. – **33**, № 2. – P. 193–207.
2. *Mani I., Bloedorn E.* Summarizing Similarities and Differences Among Related Documents // Information Retrieval. – 1999. – **1**, № 1. – P. 35–67.
3. *Radev D. R., McKeown K. R.* Generating Natural Language Summaries from Multiple Online Sources // Com-putational Linguistics. – 1998. – **24**, № 3. – P. 469–500.
4. *Multidocument* Summarization by Visualizing Topical Content / Ando R.K. et al. // Proc. ANLP/NAACL 2000 Workshop on Automatic Summarization. – 2000. – P. 79–88.
5. *Грезнева О. Ю.* Научные школы (педагогический аспект). – М.: РАО, 2003. – 69 с.
6. *Гузевич Д. Ю.* Научная школа как форма деятельности // Вопросы истории естествознания и техники. – 2003. – № 1. – С. 64–93.

Національний університет "Львівська політехніка"

Одержано 24.06.2013