Міністерство освіти і науки України
Національний університет «Львівська політехніка»

Нога Роман Юрійович

УДК 004.9:371.261

МЕТОДИ ТА ЗАСОБИ АНАЛІЗУ ТЕКСТІВ ПУБЛІКАЦІЙ ДЛЯ ДОСЛІДЖЕННЯ ДІЯЛЬНОСТІ НАУКОВИХ ШКІЛ

10.02.21 — Структурна, прикладна та математична лінгвістика

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Львів – 2015

Дисертацією є рукопис.

Робота виконана в Національному університеті "Львівська політехніка" Міністерства освіти і науки України.

Науковий керівник доктор технічних наук, професор
Шаховська Наталія Богданівна,
Національний університет "Львівська політехніка",
професор кафедри "Інформаційні системи
та мережі".

Офіційні опоненти: доктор технічних наук, професор
Воробель Роман Антонович,
Фізико-механічний інститут ім. Г. В. Карпенка НАН України, завідувач відділу обчислювальних методів і систем перетворення інформації,

кандидат технічних наук
Любченко Тетяна Петрівна,
науковий співробітник Українського мовно-інформаційного фонду НАН України

Захист відбудеться 02.07.2015 р. о 09 годині на засіданні спеціалізованої вченої ради Д 35.052.05 у Національному університеті "Львівська політехніка" (79013, м. Львів, вул. С.Бандери, 12).

З дисертацією можна ознайомитись у науково-технічній бібліотеці Національного університету "Львівська політехніка" (79013, м. Львів, вул. Професорська, 1).

Автореферат розісланий "___" _____ 2015 р.

Учений секретар
спеціалізованої вченої ради,
доктор технічних наук, професор Р.А.Бунь

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Повноцінне й оперативне забезпечення суспільства новітньою інформацією є необхідною передумовою підвищення ефективності та інноваційної віддачі наукових досліджень. Для оцінювання результативності наукової діяльності важливе місце відводиться наукометрії – напряму досліджень, що вивчає когнітивні комунікації в науці за частотою цитувань наукових робіт та їхніх авторів. Одним з елементів досліджень наукометрії є наукова школа та результати її функціонування.

Наукова школа – неформальний творчий колектив дослідників різних поколінь, об'єднаних загальною програмою та стилем дослідницької роботи, які діють під керівництвом визнаного лідера. У діяльності наукової школи реалізуються виробництво наукових знань, поширення знань, підготовку обдарованих фахівців. Одним із способів представлення результату виробництва наукових знань є наукова публікація, подана у вигляді слабоструктурованого або неструктурованого тексту. Наявність наукових шкіл є одним із визначальних факторів розвитку регіону, оскільки це напряму вказує на наявність інноваційної діяльності, а також дає змогу спрогнозувати, яку сферу діяльності доцільно розвивати. Проте велика кількість публікацій та зменшення уваги до науки та її розвитку в останні роки значно ускладнила процедуру виділення наукових шкіл та аналізу їх діяльності.

Для опрацювання електронних варіантів текстів використовуються різні методи пошуку, рубрикації чи кластеризації. Необхідно відзначити роботи Широкова В.А. в галузі математичної та прикладної лінгвістики й лексикографії, Кугурцева А.Б. з екстракції значущих ознак, Данилюка І.Г. з рубрикування текстів. До 1980-х років основними методами кластеризації текстів були експертні методи, засновані на використанні експертних оцінок для визначення тематики документів. Сьогодні цей підхід продовжує залишатися ефективним для вирішення завдань, що вимагають прийняття нетривіальних рішень про віднесення тих чи інших текстів до одного кластеру. Однак, разом з тим, ручні методи кластеризації зазвичай застосовуються тільки для невеликих колекцій текстів, а також є достатньо повільними, оскільки вимагають експертної оцінки.

Зазначені особливості методів ручної кластеризації масивів текстів зробили актуальним розроблення напівавтоматичних, а пізніше і автоматичних методів текстової кластеризації (роботи Лейкера Л., Ерк К.). Для текстів використовують такі алгоритми як Expectation maximization, Fuzzy Codok, k-середніх та інші. Проте основна проблема алгоритмів текстової кластеризації полягає у визначенні міри близькості текстів та високій обчислювальній складності.

Актуальність теми дослідження зумовлена такими факторами:

- популярність міждисциплінарних досліджень, що ускладнює віднесення публікації до наукової школи, і водночас, відсутність однозначного означення цього терміну;

- динамічність науки, швидке старіння інформації (за даними Digital Universe Study швидкість старіння інформації за 5 останніх років збільшилася вдвічі);
- велика кількість публікацій в Інтернеті ускладнює виділення основоположників та учасників наукової школи існуючими методами кластеризації, що унеможливлює налагодження зв'язків між дослідниками.

Отже, задача розроблення методів та засобів аналізу текстів публікацій для виявлення наукових шкіл та дослідження їх діяльності є актуальною.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась в рамках пріоритетного наукового напрямку, затвердженого в числі актуальних проблем Міністерством освіти і науки України, "Розроблення базових компонентів для синтезу інтелектуальних мобільних робототехнічних систем" № держреєстрації 0113U003191 (автор розробив метод кластеризації текстів наукових досліджень та метод екстракції з них даних для подальшого аналізу; зокрема, були проаналізовані тексти технічних завдань та виявлено їх міру близькості).

Мета і задачі дослідження. Метою дисертаційної роботи є розроблення є методів і засобів аналізу текстів наукових публікацій для виявлення та дослідження результатів діяльності наукових шкіл.

Мета дисертаційної роботи визначає необхідність розв'язання таких задач:

1) проаналізувати методи визначення складових публікацій та кластеризації неструктурованих даних;
2) розробити модель наукової школи та метод її виявлення;
3) розробити метод віднесення публікацій до проблематики, що розглядається науковою школою;
4) розробити метод тематичного моделювання наукових публікацій за науковими школами для прогнозування їх розвитку;
5) розробити мовно-інформаційну систему виявлення та аналізу результатів функціонування наукових шкіл та апробувати результати наукового дослідження.

Об'єктом дослідження є процес аналізу електронних версій текстів наукових публікацій та виявлення наукових шкіл.

Предметом дослідження є методи та засоби екстракції даних з текстів, кластеризації неструктурованих даних.

Методи дослідження. Для досягнення поставленої мети використано: методи кластеризації; тематичні моделі; методи прогнозування на основі часових рядів; методи об'єктно-орієнтованого аналізу та проектування – для розроблення архітектури системи.

Наукова новизна одержаних результатів. Наукова новизна роботи полягає у розв'язанні актуального завдання розроблення методів та засобів

аналізу текстів наукових публікацій для формування та дослідження функціонування наукових шкіл.. Отримано такі результати:

- *вперше* розроблено спосіб виявлення наукової школи на основі екстракції значущих ознак у слабоструктурованих документах, що стало основою для розроблення методу та засобів кластеризації текстів наукових публікацій;
- *удосконалено* метод кластеризації k-середніх електронних версій текстів наукових публікацій шляхом введення сильного зв'язку публікацій, що дало змогу здійснювати аналіз різних частин документа та враховувати їх вагомість;
- *одержав подальший розвиток* метод тематичного моделювання виявлення та дослідження результатів функціонування наукової школи на основі аналізу ймовірності появи наукових публікацій, що дало змогу визначити актуальність наукової тематики.

Практичне значення одержаних результатів. Розроблено алгоритми екстракції даних з текстів наукової публікації та їх аналізу з метою виявлення наукових шкіл. Отримано якість кластеризації на 6% вищу за інші алгоритми кластеризації (k-середніх та ієрархічної кластеризації). Розроблено алгоритм тематичного моделювання наукової школи. Це дало змогу визначити коефіцієнт кореляції між приростом статей у наукових школах за роками та кількістю захистів кандидатських та докторських дисертацій рівним 0,6. Розроблено алгоритм класифікації наукових публікацій за науковими школами як модифікацію ймовірнісних класифікаторів із визначенням релевантності публікації до наукової школи. Спроектовано архітектуру мовно-інформаційної системи для роботи з науковими публікаціями та науковими школами.

Одержані у роботі результати використано під час розроблення: інформаційної системи збору та опрацювання наукових публікацій "Рубрикатор", впровадженої в Інституті регіональних досліджень Національної академії наук України та Фізико-механічному інституті ім.Г.В.Карпенка Національної академії наук України; інформаційної системи повнотекстового пошуку, елементи якої використано в проектах компанії Рітек (для аналізу текстів технічних завдань та виявлення міри їх близькості), що підтверджено відповідними актами впровадження. Розроблення впроваджені також у навчальному процесі в курсі "Інформаційна логістика" на кафедрі інформаційних систем та мереж Національного університету "Львівська політехніка".

Особистий внесок здобувача. Усі наукові результати, подані у дисертації, одержані здобувачем особисто. У друкованих працях, опублікованих у співавторстві, внесок здобувача такий: [2] – розроблено архітектуру системи аналізу та функціонування наукових шкіл; [3, 12] – удосконалено метод кластеризації k-середніх для кластеризації текстів наукових публікацій; [4, 8, 10] – удосконалено метод прогнозування розвитку наукової школи; [14] – визначено задачі дослідження; [5, 11] –

проаналізовано методи аналізу текстової інформації; [6] – розроблено мовно-інформаційної архітектуру системи визначення наукової школи; [7, 9] – розроблено засоби формування дайджестів.

Апробація результатів дисертації. Основні результати дисертаційної роботи доповідалися на семінарах та конференціях: II Міжнародній конференції "Обчислювальний інтелект" ОІ-2013 (Черкаси, 2013); IV Міжнародній конференції "Інформаційні управляючі системи та технології" ІУСТ (Одеса, 2014); Міжнародних конференціях "The experience of designing and application of CAD systems in microelectronics" CADSM (Львів-Поляна, 2007, 2013); Міжнародній конференції "Радіоелектроніка, інформатика, управління" (Харків, 2014); VII Міжнародній науково-практичній конференції "Сучасні проблеми і досягнення в галузі радіотехніки, телекомунікацій та інформаційних технологій" (Запоріжжя, 2014); VIII Міжнародній конференції "Perspective Technologies and Methods in MEMS Design" MEMSTECH 2012 (Поляна, 2012).

Публікації. Основні результати роботи відображені у 14 опублікованих працях, у тому числі 7 статей у фахових наукових виданнях, з них 2 у закордонних журналах, що входять до наукометричних баз даних, 7 – в збірниках праць конференцій.

Структура і обсяг роботи. Дисертаційна робота складається з вступу, чотирьох розділів, висновків та додатків. Має загальний обсяг 175 сторінок, основна частина – 138 сторінок, містить 34 рисунки та 11 таблиць, 132 найменування у списку використаних літературних джерел. У додатках наведено: акти впровадження, програмні коди розробленої мовно-інформаційної системи.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету і задачі дослідження, визначено наукову новизну та практичне значення отриманих результатів, показано зв'язок роботи з науковими темами. Подано відомості про апробацію результатів роботи, публікації та особистий внесок здобувача.

У **першому розділі** проаналізовано методи опрацювання текстової інформації з множини розрізнених інформаційних ресурсів та погруповано за такими задачами: виділення основних елементів тексту; кластеризація та класифікація текстів; пошук у повнотекстових базах даних; забезпечення релевантності запиту; зменшення обсягів текстової інформації та узагальнення тексту з кількох джерел (побудова множинного реферату або дайжесту).

Проаналізовано методи та програмні продукти кластеризації, класифікації та екстракції значущих елементів з публікацій. Зокрема, немає засобів, які б дозволяли вирішувати усі задачі одночасно. Також проблемою сучасних систем аналізу текстів є те, що вони зазвичай орієнтуються на тексти англійською мовою, у той час як у роботі поставлено задачу роботи з

україномовними текстами. Такі фактори ускладнюють розв'язання задачі опрацювання тексту та роботи з ними. Особливо це актуально для наукових установ та центрів регіональних досліджень, завданнями яких є не тільки пошук та екстракція інформації з повнотекстових баз даних, але й аналіз та забезпечення підтримки прийняття рішень. Встановлено, що існує кілька визначень поняття "наукова школа", визначено їх спільні та відмінні характеристики. Показано, що кількісним та якісним вимірами результатів діяльності наукової школи є наукові публікації та захисти дисертацій.

Результати аналізу предметної області дали змогу сформулювати мету та задачі дослідження.

У **другому розділі** здійснено формальну постановку задачі формування наукових шкіл на основі аналізу наукових публікацій. Розроблено метод виділення значущих ознак з наукових публікацій. Це дало змогу здійснювати екстракцію інформації з наукових статей та уможливило подальшу кластеризацію текстових документів. Розроблено алгоритм первинної рубрикації документів, який використовується для оцінювання кількості кластерів для методу кластеризації. Розроблено метод визначення міри близькості електронних документів. Це дало змогу визначати подібність документів. Удосконалено метод k-середніх для кластеризації наукових статей. Введено означення сильного зв'язку між публікаціями. Розроблено критерій визначення якості кластеризації, що дало змогу оцінити отримане розбиття наукових публікацій на наукові школи. Розроблено метод визначення спільних ознак у назві публікації.

Уведено ряд означень.

Означення 1. *Наукова школа S* характеризується множиною наукових публікацій *Sch* з наукового напряму, визначеного через множину ключових слів *Key*, множиною авторів *Author* та множиною основоположників школи *Main*:

$$S = \langle Main, Sch, Rate \rangle, Main \in Author,$$

$$Sch_i = \langle Key, Author, T_i, Publish_i, IFactor_i, Type_i \rangle,$$

$$Author_i = \langle Surname_i, Name_i, Degree_i, Organization_i, Post_i \rangle,$$

$$Rate = f_1(.IFactor, \Delta Degree),\ f_2 : T_i \to Key.$$

Публікація Sch_i характеризується множинами ключових слів Key та авторів $Author$, повним текстом T_i, виданням $Publish_i$, рейтингом $IFactor_i$ та типом $Type_i$. Автор $Author_i$, відповідно, – кортеж з таких характеристик, як прізвище, ім'я, науковий ступінь, організація, посада; $Rate = f_1(IFactor, \Delta Degree)$ – функція визначення рейтингу наукової школи на основі індексу публікацій та приросту кількості авторів з науковими ступенями; $IFactor$ – показник цитованості журналів, визначає інформаційну значимість наукових журналів:

$$Rate = \frac{1}{m}\sum_{i}^{n} k_i IFactor_i \Delta Degree,$$

де $m = |Sch|$ – кількість публікацій у науковій школі, k_i – кількість публікацій представників шкіл у виданні з рейтингом $IFactor_i$.

$$Degree_{t_i} = 100 Deg(Doctor) + Deg(Cand),$$

$$\Delta Degree = \frac{Degree_{tyear} - Degree_{t_0}}{year},$$

де $Degree_{t_i}$ – кількість кандидатів $Deg(Cand)$ та докторів наук $Deg(Doctor)$ на t_i році спостереження за розвитком школи, $year$ – кількість років спостереження за школою. Цей показник дає змогу забезпечити аналіз наявності ланки "учитель – учень" у науковій школі. Визначено функцію f_2 отримання ключових слів на основі аналізу тексту T_i наукової публікації: $f_2 : T_i \to Key$. На основі цієї функції побудовано метод екстракції ключових слів з текстів наукових публікацій.

Вхідною інформацією для віднесення публікації до наукової школи є текстовий файл будь-якого формату з вмістом публікації. З файлу необхідно визначити базові елементи публікації: Автор(и) публікації; Наукова установа; Назва публікації; Ключові слова; Основний текст.

Наукові публікації є слабоструктурованими електронними документами (ЕД). Визначено елементи ЕД, які отримані на основі повнотекстового пошуку та екстракції. Наукові публікації можуть мати різний тип – статті, тези конференції, автореферати, дисертації, книги, звіти тощо. Елемент ЕД "Основний текст" також має внутрішню структуру, елементи якої розділені заголовками.

Розроблено метод виділення (екстракції) складових елементів наукової публікації з їх подальшим аналізом (функція f_2). Наукова публікація, яка належить до певної наукової школи, складається з послідовності речень $A_1, A_2, ..., A_l$ та утворює кортеж $T = (A_1, A_2, ..., A_l)$, а речення $A_i, i = \overline{1,l}$ – з послідовності слів $a_{ij}, i = \overline{1,l}, j = \overline{1,n}$, яке, у свою чергу, зображується кортежем $r_i = (a_{i1}, a_{i2}, ..., a_{il})$. Зміст (семантику) тексту T позначено $S(T)$. Визначено множину ключових слів $Key = \{key_1, key_2, ..., key_m\}$ наукової школи, виділених з досліджуваного тексту: у реченні $r = (a_1, a_2, ..., a_l)$ знаходять ключове слово $a_p (a_p \in Key)$. Слова, довжина яких є не більшою ніж три літери, виконують у тексті службову роль і не впливають істотно на семантику речення. Необхідні кроки для виділення з контенту необхідної інформації для подальшої роботи з нею: завантаження ЕД, реферування ЕД, екстракції елементів.

Метод екстракції значущих характеристик базується на понятті ваги речення та ваги слова (словосполучення). Основу аналітичного етапу складає процедура призначення вагових коефіцієнтів для кожного блоку тексту відповідно до таких характеристик, як розташування цього блоку в оригіналі, частота появи в тексті, частота використання в ключових реченнях, показники статистичної значущості.

Сума індивідуальних ваг слів та речення, визначена після додаткової модифікації відповідно до спеціальних параметрів налаштування, пов'язаних з кожною вагою, формує загальну вагу речення *U*:

$$Weight(U) = WordsWeighn(U) + 10 * Place(U) + 10 * Format(U). \quad (1)$$

Основна частина, у свою чергу, ділиться на фрагменти за підрозділами та розділами, введеними авторами. Більша вага надається коефіцієнтам розташування та форматування, аніж вазі слова.

Коефіцієнт розташування визначено як:

$$Place(U) = \begin{cases} 0, \left(ns / n_{count} > 0{,}9 \right) \vee \left(ns / n_{count} < 0{,}1 \right) \\ 1, \left(0{,}1 \le ns / n_{count} < 0{,}3 \right) \vee \left(0{,}7 < ns / n_{count} <= 0{,}9 \right) \\ 2, \left(0{,}3 \le ns / n_{count} \le 0{,}7 \right) \end{cases}, \quad (2)$$

де ns – номер речення, а n_{count} – загальна кількість речень у документі. Початок та кінець тексту оцінюються меншим значенням (бо це переважно вступ та висновок) 0-1, а середина – 2. Також, якщо у документі є анотація, то цьому фрагменту тексту присвоюється $Place(U) = 4$.

Коефіцієнт форматування речення *U* визначається як:

$$Format(U) = \begin{cases} 0, \text{вирівнювання зліва або справа}, \\ 1, \text{вирівнювання по ширині}, \\ 2, \text{вирівнювання по центру}. \end{cases} \quad (3)$$

Коефіцієнт $WordsWeight(U)$ визначається як середня вага слова у реченні (сума ваг усіх ключових слів $Weight(Q)$, що входять до речення, поділена на кількість ключових слів у реченні).

Вага терміну *Q* визначається за формулою:

$$Weight(Q) = Frequency(Q) + Place(Q) + Format(Q) + User(Q). \quad (4)$$

Частотний коефіцієнт $Frequency(Q)$ – відношення кількості входження деякого слова (*word*) до загальної кількості слів (*words*) документа. Таким чином, оцінюється важливість слова в межах окремого документа:

$$Frequency(Q) = \frac{word}{words}. \quad (5)$$

Коефіцієнт розташування $Place(Q)$ визначається як функція належності до речення, де зустрічається слово однієї з ключових фраз:

“Ключові слова”, “Key words”. Якщо така фраза зустрілась, то коефіцієнт розташування рівний 5, якщо ні – то 0.

Коефіцієнт форматування слова $Format(Q)$ визначається залежно від того, чи слово виділене жирним, курсивом чи підкреслене. Якщо слово зовсім не відформатоване, то коефіцієнт дорівнює 0; якщо одним форматом, то – 1; якщо двома, то – 2; якщо трьома, то – 3.

Показник $User(Q)$ формується на основі оцінювання слова користувачем ($User(Q) \in [0..10]$).

Вагові коефіцієнти, використані у формулі (1), отримані емпірично. У роботі ставилася задача не точного визначення їх значень, а встановлення ваги певних адитивних параметрів. Тому для цих коефіцієнтів важливим є порядок числа, а не його значення.

Результатом методу виділення складових текстового документа є вектор, у якому для таких характеристик як автор, наукова установа використовуються бінарні ознаки, а для ключових слів – ваги.

Модифікований метод k-середніх полягає у виконанні таких кроків.

1.Задаємо кількість кластерів *k*, $N \geq k \geq 2$, де *N* – кількість публікацій.

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

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

$$d(X,Y) = \sum_i^p l(X.A_i, Y.A_i) + \sum_i^r l(X.D_i, Y.D_i) + \sum_i^w l(X.B_i, Y.B_i) + l(X.C, Y.C),$$

де функція *l* повертає 1, якщо обидва її параметри мають однакові значення, та 0 в іншому випадку, X, Y – електронні версії текстів наукових публікацій, p – кількість авторів у текстах публікацій X, Y, r – сумарна кількість ключових слів, w – сумарна кількість наукових установ, $X.A_i$ – значення автора X_i публікації X, $X.C$ – значення назви *C* наукової статті X.

2. Обираємо *k* об’єктів, які вважатимемо центрами відповідних кластерів (центроїдами). Покласти номер кроку $s = 0$.

3. Формуємо вектор центроїдів $\left\langle cx_1^s, cx_2^s, \ldots, cx_k^s \right\rangle$ (центрів ваги).

Для кожного об’єкта знаходимо відстань до усіх центроїдів. Для знаходження відстані використовуємо Евклідову метрику.

4. Шукаємо матрицю відстаней до центроїдів кластерів та формуємо кластери S_i, $i = \overline{1,k}$:

$$\min\left[\sum_{j=1}^{k}\sum_{i}^{N}\left\|x_i - cx_j\right\|^2\right],$$

де N – кількість публікацій, cx_j – центроїд кластера з номером j.

Після розрахунку матриці відстаней шукаються *сильні зв'язки* об'єкта з кластером.

Означення 2. *Сильним* названо зв'язок між об'єктами X та X_i, якщо значення відстані назв публікацій менше, ніж третина від максимальної відстані серед усіх назв публікацій:

$$d_s(X, X_i) \le \frac{\max\left(d(X, X_1),..., d(X, X_N)\right)}{3}.$$

5. Шукаємо вартість розбиття:

$$Cost = \sum_{i=1}^{k} \sum_{j=1}^{|S_i|} d_{ij} d_s(x_j, cx_i),$$

де k – кількість кластерів, $|S_i|$ – кількість об'єктів у кластері S_i, d_{ij} – відстань до центру кластера i.

6. Шукаємо нові центроїди кластерів:

$$cx_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j .$$

Якщо $\|CX^s\| \neq \|CX\|$, то $s = s + 1$. Перейти на крок 3.

7. Якщо *Cost* не задовольняє умовам локального оптимуму, k=k+1 і перейти на крок 3.

Для випадків, коли в кластері знайдено сильні зв'язки, застосовано метод визначення спільних ознак у назві публікації. Для цього введено поняття міри відстані назв. Ознаки у назві статей називатимемо термами.

Нехай в колекції Q n-загальна кількість термів у всіх документах, n_i – кількість термів в документах, в яких зустрічається терм i. Нехай загальне число термів j у всіх текстах – N_j, а кількість термів j в документах, що містять терм i – N_{ij}. Тоді величина

$$\rho_{ij} = \frac{\left(\frac{n_i}{n}\right)^{N_j} \left(1 - \frac{n_i}{n}\right)^{N_j - N_{ij}} N_j!}{\left(N_j - N_{ij}\right)!}$$

є мірою кореляції між термами i і j – чим вона менша, тим більше корельовані ці терми. Тоді сила зв'язку термів i і j $\rho = \max\left(\rho_{ij}, \rho_{ji}\right)$ слугує мірою кореляції термів i і j у випадку $\rho_{ij} \neq \rho_{ji}$.

Терм t, що міститься у текстовій колекції Q, називається *значущим* (характерним) на рівні β, якщо різниця частоти, з якою терм t зустрічається в колекції Q і середньої частоти, з якою він зустрічається на множині наукових публікацій, перевищує β.

Якість кластеризації визначено як нормоване значення кількості правильно та неправильно віднесених та правильно та неправильно відкинутих публікацій до кластера.

У **третьому розділі** розроблено метод визначення ймовірності появи нових публікацій у наукових школах. Побудовано алгоритм пошуку публікацій за науковими школами, а також пошуку спорідненої наукової школи. Розроблено алгоритм класифікації публікацій за відомими науковими школами (рубриками).

Розроблено *метод тематичного моделювання наукових публікацій за науковими школами*. Аналізуються не усі слова ЕД (наукової публікації), а лише ключові. Тоді ймовірнісна модель появи пари "школа-ключове слово" подана як:

$$p(d, key) = \sum_{s \in S} p(s)\, p(key \mid s)\, p(d \mid s) = \sum_{s \in S} p(s)\, p(key \mid s)\, p(s \mid d) =$$

$$= \sum_{s \in S} p(key)\, p(s \mid key)\, p(d \mid s),$$

де S – множина шкіл; $p(s)$ – невідомий апріорний розподіл шкіл у всій колекції; $p(d)$ – апріорний розподіл на множині наукових публікацій, емпірична оцінка $p(d) = n_d / n$, де $n = \sum_d n_d$ – сумарна довжина всіх публікацій; $p(key)$ – апріорний розподіл на множині ключових слів, емпірична оцінка $p(key) = n_{key} / n$, де n_{key} – число входжень ключового слова *key* у всі публікації.

Множина наукових публікацій містить для кожної публікації d додаткову інформацію, яку називають метаінформацією:

- список авторів публікації A;
- список публікацій d', на які посилається d;
- список авторів A, на яких посилається d;
- список публікацій, які посилаються на d;
- список авторів, які посилаються на d;
- список наукових шкіл, до яких відноситься d.

Шукані ймовірності розподілу $p(key|s)$, $p(s|d)$ виражено через $p(s|key)$, $p(d|s)$ за формулою Байєса:

$$p(key \mid s) = \frac{p(s \mid key)\, p(key)}{\sum_{w'} p(s \mid key')\, p(key')}; \; p(s \mid d) = \frac{p(d \mid s)\, p(s)}{\sum_{s'} p(d \mid s')\, p(s')},$$

де key', s' – список ключових слів та наукова школа відповідно, отримані з публікацій, на які посилається d.

Для ідентифікації параметрів тематичної моделі (школи) за колекцією наукових публікацій застосовано принцип максимуму правдоподібності, який зведено до задачі мінімізації:

$$\sum_{d\in D}\sum_{key\in d} n_{d,key}\log p(d,key)\to\min;\ \sum_{key} p(key\mid s)=1,\sum_{s} p(s\mid d)=1,\sum_{s} p(s)=1,$$

де $n_{d,key}$ – число входжень ключового слова *key* в публікацію *d*.

Для розв'язання цієї оптимізаційної задачі застосовано ЕМ-алгоритм.

Прогнозування зміни динаміки публікацій здійснено за допомогою часових рядів, а саме методом ковзаючого середнього. Задачею прогнозування є знаходження залежності між кількістю публікацій по кожній із знайдених наукових шкіл, частотою появи нових ключових слів та частотою отримання наукових ступенів представниками шкіл. Динаміка зміни кількості ключових слів охарактеризована стосовно базисного спостереження і величини зміни сусідніх рівнів. В якості статистичних характеристик часового ряду Y_i, $i=\overline{1,n}$ використано середнє арифметичне кількості публікацій $Y=\frac{1}{N}\sum_{j=1}^{N}Y_j$ та середній абсолютний приріст кількості публікацій за школами $\overline{Y}=(Y_n-Y_1)/(N-1)$, де N – кількість рівнів ряду, Y_i – рівні ряду. Відповідно до методу перевірки істинності різниці середніх початковий часовий ряд розбивається на дві однакові частини, після чого перевіряється гіпотеза про істотність різниці середніх для цих частин. Перевірка однорідності даних здійснено на основі критерію Ірвіна, що заснований на порівнянні сусідніх значень ряду. Відповідно до нього розраховується характеристика $ts=\frac{Y_t-Y_{t-1}}{Y}$.

Аналіз автокореляції виконано за допомогою графіка і критичних значень коефіцієнтів, встановлених експертно. Параметри цього рівняння знаходять за методом найменших квадратів. Ковзну середню в обраному інтервалі визначено як зважене середнє усіх попередніх рівнів. Метод найменших квадратів використано також для пошуку залежності між приростом кількості публікацій у наукових школах за роками та приростом кількості захистів дисертаційних робіт $\Delta Degree$.

З цією метою здійснено завантаження файлів з сайту МОН України (додатки). Структура файлів визначається сталим форматуванням та складається з таких компонентів:

1) науковий ступінь (доктор, кандидат),
2) галузь знань,
3) навчальний заклад (наукова установа),
4) спеціалізована вчена рада (не враховується),
5) прізвище, ім'я, по батькові (ПІП), спеціальність (остання характеристика не враховується).

Далі здійснено пошук виділених ПІП у базі даних авторів та визначення залежності між кількістю публікацій та кількістю захистів.

Алгоритм пошуку залежності між приростом кількості публікацій у наукових школах за роками та приростом кількості захистів дисертаційних робіт подано на рис. 1.

Рисунок .1. Схема алгоритму пошуку залежності між приростом кількості публікацій у наукових школах за роками та приростом кількості захистів дисертаційних робіт

Далі у розділі розроблено *метод класифікації наукових публікацій***.**

Для цього визначено релевантність документа певному класові (науковій школі).

Крок 1. Нормалізація

Нормалізація – спосіб зменшення абсолютного значення ваги індексних термінів, виявлених у ЕД. Обрано косинусну нормалізацію. За використання цього методу нормалізації вага кожного індексного терміна ділиться на Евклідову довжину вектора документа, що оцінюється. Евклідова довжина вектора визначається як:

$$L = \sqrt{w_1^{\,2} + w_2^{\,2} + \ldots + w_{n^2}} \,,$$

де $w_i = Weight(Q)$ – вага i-го терміна Q у документі.

Остаточна формула для обчислення ваги w терміна Q у документі з урахуванням косинусного фактора нормалізації подана формулою:

$$W = \frac{Weight(Q)}{\sqrt{w_1^2 + w_2^2 + \ldots + w_n^2}}.$$

Терміни, що відсутні в тексті документа, мають нульову вагу. У списку, що повертається як результат виконання запиту, документи впорядковуються за зменшенням цього чисельного значення.

Крок 2. Розрахунок умовних ймовірностей.

Для представлення наукових публікацій використовується векторна модель, де будь-який документ характеризується вектором $x = x_1, x_2, \ldots, x_n$, де значення характеристик Автор, Наукова установа рівні x_i = 0 або 1 залежно від того, чи є присутнім у тексті i-й індексний термін чи ні, i належить діапазонові від 0 до 1 для характеристики Ключові слова.

Розглядаються дві взаємно виключаючі події:

- w_1 – документ відноситься до наукової школи x_i;
- w_2 – документ не відноситься до наукової школи x_i.

Для кожної наукової публікації обчислено умовні ймовірності $P(w_1 \mid x_i)$ і $P(w_2 \mid x_i)$ для визначення, які документи відносяться до наукової школи, а які ні:

$$P(w_i \mid x) = \frac{P(x \mid w_i)P(w_i)}{P(x)}, i = 1,2 \cdot$$

У наведеній формулі $P(w_1)$ – первісна ймовірність відповідності (i=1) або невідповідності (i= 2) запитові, величина $P(x \mid w_i)$ пропорційна ймовірності відповідності або невідповідності науковій школі для заданого x; у недискретному випадку вона являє собою функцію щільності розподілу й позначається як $P(x \mid w_i)$.

Крок 3. Визначення ймовірності віднесення до класу:

$$P(x) = \sum_{i=1}^{2} P(x \mid w_i)P(w_i),$$

що являє собою імовірність віднесення наукової публікації x до певного класу. Величина $P(x)$ виступає як фактор, що нормалізує (тобто з її допомогою досягається виконання умови $P(w_1 \mid x) + P(w_2 \mid x) = 1$).

Для визначення релевантності документа певній науковій школі використано правило: якщо $P(w_1 \mid x_i) > P(w_2 \mid x_i)$, то наукова публікація належить до наукової школи x_i. Для множини наукових шкіл визначатимемо вектор значень $P(w_1 \mid x_i)$.

У розділі розроблено метод тематичного моделювання публікацій наукової школи, алгоритм пошуку наукових статей за параметрами користувача, алгоритм класифікації наукових публікацій за відомими науковими школами (рубриками).

У **четвертому розділі** розроблено мовно-інформаційну систему кластеризації наукових публікацій, що уможливило подальше дослідження розроблених методів на предмет доцільності їх використання для інших задач. Побудовано архітектуру, схему бази даних та основні програмні модулі. Апробовано результати роботи розроблених алгоритмів для задач: рубрикування публікацій, що виникає у електронних бібліотеках чи наукових установах; кластеризації технічної документації. Визначено якість рубрикування та встановлено залежність між величинами помилок першого та другого роду та обсягом вибірки. Кореляційний момент між обсягом вибірки в класі та кількістю помилок першого роду становить 0,76. Апробовано результати роботи методів визначення перспективності наукової школи. З цією метою встановлено залежність між кількістю статей у науковій школі за роками та кількістю захистів. Проаналізовано результати

кластеризації відомими методами. Встановлено, що відсоток правильно прокластеризованих документів та документів, що правильно не потрапили у клас, є найвищий для розробленого методу і становить 94% на спільній для усіх методів колекції наукових статей. Якість кластеризації залежить від кількості ключових слів і рівня їх перетину в рубриках, а також від наявності статей з авторами, що належать до різних наукових шкіл.

Програма складається з наступних модулів: База даних; Підсистема графічного представлення; Підсистема кластеризації наукових статей за науковими школами; Підсистема визначення вагомості та швидкості росту наукової школи (рис. 2).

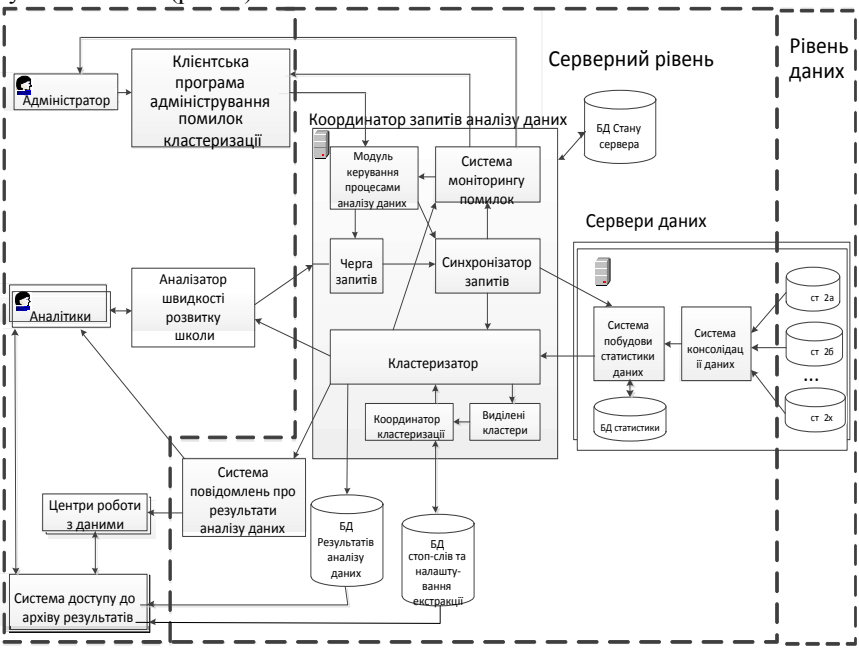

Рисунок 2. Архітектура мовно-інформаційної системи кластеризації наукових публікацій

Для тестування роботи системи опрацьовано 134 файли наукових публікацій. «Правильна» рубрика текстових документів відома наперед та встановлена експертно. Проаналізовано якість рубрикації (*TP* (*true positive*) – кількість ЕД, правильно віднесених до категорії; *FP* (*false positive*) – помилка другого роду – кількість ЕД, неправильно віднесених до категорії; *FN* (*false negative*) – помилка першого роду – кількість ЕД, які неправильно відкинуті; *TN* (*true negative*) – кількість ЕД, які правильно відкинуті) (табл. 1).

Середнє нормоване значення правильно рубрикованих документів становить 94 %. Проаналізовано залежність якості кластеризації від обсягу публікацій у рубриці. Чим більшою є «загальність» рубрики, тим важче її кластеризувати (рис. 3).

Таблиця 1.

Результати аналізу якості кластеризації

Клас	nTP	nFP	nFN	nTN
database	93%	11%	7%	33%
computer science	93%	13%	7%	25%
programming	96%	2%	4%	50%
network	94%	6%	6%	60%
system analysis	93%	7%	7%	50%

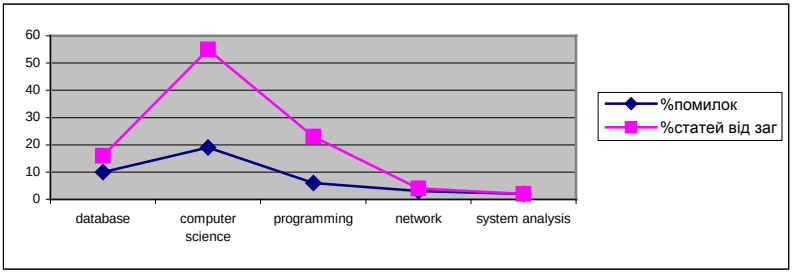

Рисунок 3. Залежність помилки першого роду від обсягу вибірки

Проаналізовано якість кластеризації залежно від кількості ключових слів, а також від ступеню їх перетину. Алгоритм тестувався на чотирьох колекціях вхідних даних з однаковою кількістю об'єктів у кожному з класів, але з різною кількістю ключових слів та з різною кількістю спільних для класів ключових слів. Результати аналізу подано у табл. 2.

Таблиця 2.

Залежність якості кластеризації від кількості ключових слів

	Колекція 1		Колекція 2	
Клас	к-сть ключових слів	nTP	к-сть ключових слів	nTP
database	7	87	16	88
computer science	11	67	26	62
programming	12	69	19	67
network	3	93	7	91
system analysis	4	94	5	89

Визначено якість кластеризації для різних методів. Для порівняння проаналізовано результати роботи трьох інших алгоритмів на тих же колекціях (табл. 3). Таким чином, розроблений алгоритм продемонстрував кращі результати для величини nTp на текстових колекціях у порівнянні з іншими розглянутими алгоритмами.

Таблиця 3.

Порівняння результатів роботи різних методів кластеризації

Метод кластеризації	nTP, %
Розроблений метод кластеризація	92
Острівна кластеризація	86
К-середніх	71
Average Link	78

Далі проаналізовано виділені кластери на предмет того, чи дійсно вони є науковими школами. З цією метою порівнювались множини публікацій, сформовані розробленим методом на основі аналізу текстів та їх кластеризації, та публікації, опубліковані науковцями офіційно визнаних наукових шкіл. «Правильність» кластерів відома і оцінена так, як і якість рубрикування. На відміну від рубрикування, під час кластеризації враховуються також відомості про авторів публікацій. Аналізувались статті з авторами, що належали до різних наукових шкіл (табл. 4).

Таблиця 4.

Залежність % помилок кластеризації від % авторів з різних наукових шкіл

% публікацій з авторами з різних наукових шкіл	%помилок
0	3
4	12
9	19
18	27

Для визначення перспективності школи протягом 3-х років аналізувалися файли з інформацією про захисти кандидатських та докторських дисертацій. Проаналізовано ймовірність появи нових публікацій у виділених наукових школах залежно від різних параметрів. Здійснено передбачення появи нових публікацій за школами (рис. 4).

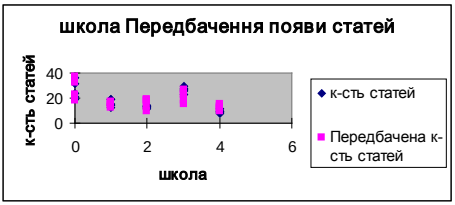

Рисунок 4. Передбачення появи публікацій за школами

У розділі розроблено мовно-інформаційну систему кластеризації наукових публікацій, апробовано результати роботи розробленої системи. Визначено якість рубрикування та встановлено залежність між величинами помилок першого та другого роду та обсягом вибірки рубрики.

У **додатках** наведено акти впровадження результатів дисертаційної роботи та програмний код розробленої мовно-інформаційної системи.

ВИСНОВКИ

У роботі розв’язано актуальне наукове завдання розроблення математичних методів і програмних засобів аналізу електронних версій текстів наукових публікацій для виявлення та дослідження результатів функціонування наукових шкіл, що дає змогу підвищити якість прийняття рішень щодо доцільності підтримки наукових досліджень.

У результаті виконання цієї роботи одержано такі результати:

1. Проаналізовано сучасні тенденції розвитку інформаційних технологій для аналізу слабоструктурованих даних. Виявлено нерозв'язані задачі в галузі екстракції інформації з електронних версій документів та їх кластеризації.

2. Наукову школу визначено через множину наукових публікацій та їх авторів. Це стало основою для розроблення методу екстракції значущих ознак зі слабоструктурованих документів.

3. Розроблено метод виявлення наукової школи на основі екстракції значущих ознак у слабоструктурованих документах, що стало основою для розроблення методу та засобів кластеризації та класифікації текстів наукових публікацій.

4. Удосконалено метод кластеризації k-середніх текстів наукових публікацій шляхом введення сильного зв'язку публікацій, що дало змогу здійснювати аналіз різних частин документа та враховувати їх вагомість. Удосконалений метод показав кращі на 6% результати кластеризації порівняно з методами острівної та ієрархічної кластеризації, а також на відміну від немодифікованого методу k-середніх не залежить від початкового вказання кількості кластерів.

5. Визначено поняття якості рубрикування та встановлено залежність між величинами помилок першого та другого роду та обсягом вибірки рубрики. Кореляційний момент між обсягом вибірки в класі та кількістю помилок першого роду удосконаленим методом кластеризації k-середніх становить 0,76.

6. Удосконалено метод тематичного моделювання публікацій наукової школи шляхом визначення ймовірності появи пар «публікація-школа» та «школа-захист», що дало змогу визначити актуальність наукової тематики, прогнозувати появу нових публікацій та встановити залежність між кількістю статей в науковій школі за роками та кількістю захистів.

7. Розроблено архітектуру мовно-інформаційної системи кластеризації наукових публікацій, що уможливило подальше дослідження розроблених методів на предмет можливості їх використання для інших задач.

СПИСОК ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Нога Р. Ю. Метод формування наукових шкіл на основі аналізу елементів публікацій / Нога Р. Ю. // Математичні машини і системи. – 2013. – №4. – С. 107-114.
2. Нога Р. Ю. Розроблення архітектури системи аналізу функціонування наукових шкіл / Нога Р. Ю., Шаховська Н. Б. // Відбір та обробка інформації. – 2013. – № 39(115). – С. 94-97.
3. Shakhovska N. One method of analysis of research publications' elements / Shakhovska N., Noha R. // MEST Journal. – 2014. – V.2, is.2. – P. 94-102. – Режим доступу: http://mest.meste.org/MEST_Najava/III_shakhovska.pdf
4. Шаховська Н. Б. Метод кластеризації наукових публікацій для формування наукової школи та прогнозування динаміки її розвитку /

Шаховська Н. Б., Нога Р. Ю. // Моделювання та інформаційні технології. – Вип. 70. – Київ, 2014. – С. 113-117.

5. Шаховська Н. Б. Аналітичний огляд методів та засобів опрацювання текстової інформації / Шаховська Н. Б., Нога Р. Ю. // Вісник Національного університету "Львівська політехніка". – Л. : Вид-во Нац. ун-ту "Львів. Політехніка", 2011. – №715: Інформаційні системи та мережі. – С. 323-331.
6. Shakhovska N. The system developing of forming research schools basis of publication elements analysis / Shakhovska N., Noha R. // Applied Computer Science Journal. – 2014. – V. 10, No. 2. – P. 57-66.
7. Шаховська Н. Б. Інтелектуальна система анотування новин для оцінювання достовірності їх джерел / Шаховська Н. Б., Нога Р. Ю., Вовк О. Б. // Моделювання та інформаційні технології. – 2013. – Вип. 70. – С. 113-117.
8. Шаховська Н. Б. Метод формування наукової школи та прогнозування динаміки її розвитку / Шаховська Н. Б., Нога Р. Ю. // Обчислювальний інтелект (результати, проблеми, перспективи): матеріали II Міжнародної науково-технічної конференції, 14-17 травня 2013 р., м. Черкаси. – Черкаси, 2013. – С. 311-313.
9. Shakhovska N. Building a smart news annotation system for further evaluation of news validity and reliability of their sources / Shakhovska N., Noha R., Vovk O. / Інформаційні управляючі системи та технології (ІУСТ - Одеса - 2014) = Information Control Systems and Technologies: матеріали Міжнар. наук.-практ. конф., 23-25 вересня 2014 р., м. Одеса. – Одеса : "Видавінформ" ОНМА, 2014. – С. 290-295.
10. Noha R. Connection between publications analysis and research schools forming / R. Noha, A. Noha, I. Garandzha / Досвід, розробка і застосування САПР в мікроелектроніці: матеріали XIII міжнар. конф., CADSM – 2013, 19-23 лютого 2013 р., м.Львів; Поляна. – Львів, 2013. – С.259-261.
11. Noha R. Method of forming scientific schools on the basis of research publications elements analysis / R. Noha, I. Garandzha / VIIIth International Conference MEMSTECH 2012 "Perspective Technologies and Methods in MEMS Design", 18-21 April 2012, Polyana-Svalyava, Ukraine. – Lviv, 2012. – P. 154.
12. Noha A. New approaches to the decision of management problem functioning energetic objects in the conditions of the destabilizing factors / Noha A., Noha R., Sikora L. / IXth International Conference on the Experience of Designing and Application of CAD Systems in Micro-electronics, February, 20-24, 2007, Polyana, Ukraine. – Polyana, 2007. – P. 374-375.
13. Нога Р. Ю. Метод формування наукових шкіл на основі аналізу елементів публікацій / Нога Р. Ю. / Материалы 16 международного молодежного форума "Радиоэлектроника и молодежь в XXI веке" , 17-19 апреля 2012. – Харьков : ХНУРЭ, 2012. – Т. 6. – С. 11-12.
14. Shakhovska N. The method of scientific paper automatic abstracting / Shakhovska N, Noha R. / Сучасні проблеми і досягнення в галузі радіотехніки, телекомунікацій та інформаційних технологій: матеріали VII міжнародної науково-практична конференції, 17 – 19 вересня 2014 р., м. Запоріжжя. – Запоріжжя, 2014. – С. 116-117.

АНОТАЦІЇ

Нога Р. Ю. Методи та засоби аналізу текстів публікацій для дослідження діяльності наукових шкіл. – На правах рукопису.

Дисертаційна робота на здобуття наукового ступеня кандидата технічних наук за спеціальністю 10.02.21 – Структурна, прикладна і математична лінгвістика. – Національний університет "Львівська політехніка" МОН України, Львів, 2015.

У дисертаційній роботі розв'язано актуальне наукове завдання розроблення математичних методів і програмних засобів аналізу текстів наукових публікацій для виявлення та дослідження результатів функціонування наукових шкіл, що дає змогу підвищити якість прийняття рішень щодо доцільності підтримки наукових досліджень за рахунок виявлення нових знань у слабоструктурованих документах.

Проаналізовано методи опрацювання текстової інформації з множини розрізнених інформаційних ресурсів та визначено можливість їх застосування до аналізу наукових публікацій. Удосконалено метод екстракції даних з наукової публікації. Розроблено алгоритм попередньої рубрикації наукових публікацій з метою визначення ймовірної кількості кластерів. Удосконалено метод кластеризації k-середніх для поділу наукових статей за науковими школами. Визначено метрику якості кластерного рішення. Розроблено алгоритми аналізу наукових публікацій та прогнозування зміни кількісних характристик наукових шкіл таких як кількість публікацій, захисти дисертацій. Спроектовано архітектуру системи аналізу текстів публікацій для формування та аналізу наукових шкіл. Апробовано розроблені методи для електронної бібліотеки та для наукової установи.

Ключові слова: наукова школа, аналіз тексту публікації, екстракція інформації, кластеризація, тематичне моделювання.

Нога Р. Ю. Методы и средства анализа текстов публикаций для исследования деятельности научных школ. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 10.02.21 – Структурная, прикладная и математическая лингвистика. – Национальный университет "Львовская политехника" МОН Украины, Львов, 2015.

В диссертационной работе решено актуальное научное задание разработки математических методов и программных средств анализа текстов научных публикаций для выявления и исследования результатов функционирования научных школ, что позволяет повысить качество принятия решений о целесообразности поддержки научных исследований за счет выявления новых знаний в слабоструктурированных документах.

Проанализированы методы обработки текстовой информации из множества разрозненных информационных ресурсов. Рассматриваются существующие методы анализа и работы с текстовыми данными, их преимущества, области применения, ограничения и проблемы.

Проанализирована возможность их применения к анализу научных публикаций. Определены элементы текстовых документов, которые должны быть получены на основе полнотекстового поиска и экстракции. Усовершенствованы методы экстракции данных из научной публикации и кластеризации k-средних для разделения научных статей по научным школами. Определена метрика качества кластерного решения.

Разработан метод определения вероятности появления новых публикаций в научных школах. Предложены алгоритмы анализа научных публикаций и прогнозирования изменения количественных характристик научных школ, таких как количества публикаций и защит диссертаций. Разработан алгоритм классификации публикаций по известным научными школами (рубрикам). Спроектирована архитектура системы анализа текстов публикаций для формирования и анализа научных школ. Определено качество кластеризации. Построено схему базы данных и основные программные модули. Они могут использоваться не только для выявления научных школ на базе анализа текстов публикаций, но и библиотеками для хранения и поиска публикаций, и центрами развития и инноваций для выявления тематик, по которым за определенный период больше публикаций.

Ключевые слова: научная школа, анализ текста публикации, экстракция информации, кластеризация, тематическое моделирование.

Noha R. Yu. Methods and tools for text analysis publications to identify and study the functioning scientific schools. – The manuscript.

The thesis for the degree of candidate of technical sciences, specialty 02.10.21 – Structural, applied and mathematical linguistics. – Lviv Polytechnic National University, Ministry of Education and Science of Ukraine, Lviv, 2015.

The dissertation solved the problem of current scientific development of mathematical methods and software tools to analyze texts to identify scientific publications and research results of functioning scientific schools, allowing to increase the quality of decision-making regarding the advisability support research through the discovery of new knowledge in semistructured documents.

There are described the methods of processing text information from a plurality of disparate information resources. The method of extraction of data from scientific publications is given. The method of k-means clustering to split research papers for academic schools. There is defined the quality metric of cluster solution. The method of determining the likelihood of new publications in scientific schools is described. There is designed system architecture development and evaluation of scientific schools are given. Developed methods were tested for e-libraries and for academic institutions.

Keywords: scientific school, publication text analysis, information extraction, clustering, thematic design.