МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Одеський національний політехнічний університет

КОЛЯДА АНДРІЙ СЕРГІЙОВИЧ

УДК 004.6:001(043.3/.5)

МОДЕЛІ І МЕТОДИ ПОШУКУ ІНФОРМАЦІЇ У НАУКОМЕТРИЧНИХ БАЗАХ ДАНИХ

05.13.06 – Інформаційні технології

Автореферат дисертації на здобуття наукового ступеня
кандидата технічних наук

Одеса – 2015

Дисертацією є рукопис.

Робота виконана в Одеському національному політехнічному університеті Міністерства освіти і науки України.

Науковий керівник:

доктор технічних наук, професор **Гогунський Віктор Дмитрович**,
Одеський національний політехнічний університет,
завідувач кафедри Управління системами безпеки життєдіяльності.

Офіційні опоненти:

доктор технічних наук, професор **Мещеряков Володимир Іванович**,
Одеський державний екологічний університет,
завідувач кафедри інформаційних технологій;

кандидат технічних наук, доцент **Палій Сергій Володимирович**
Київський національний університет будівництва та архітектури,
доцент кафедри основ інформатики, начальник Центру інформаційних технологій.

Захист відбудеться 1 жовтня 2015 р. о 13:30 на засіданні спеціалізованої вченої ради Д 41.052.01 в Одеському національному політехнічному університеті за адресою: 65044, м. Одеса, пр. Шевченка, 1, ауд. 400 – А.

З дисертацією можна ознайомитися в бібліотеці Одеського національного політехнічного університету за адресою: 65044, м. Одеса, пр. Шевченка, 1

Автореферат розісланий 28 серпня 2015 р.

Вчений секретар О.Є. Колесніков

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність дисертаційного дослідження. Розвиток інформаційних технологій з організації міжнародних наукометричних баз даних (НБД) і електронних бібліотек породжує нові можливості і завдання у сфері освітньої та наукової діяльності вищої школи України. Одним з напрямів цієї діяльності є визначення узагальненої оцінки якості і результатів наукових досліджень окремого вченого, кафедри, університету та вищих навчальних закладів України в цілому. Аналіз характеристик та основних властивостей НБД та індикаторів цитування наукових публікацій відображають широкий спектр цільового призначення НБД - від забезпечення суто інформаційних потреб науковців до всебічного аналізу публікаційної активності авторів.

На сьогоднішній день МОН України цілеспрямовано орієнтує публікаційну діяльність учених на входження у світове наукове співтовариство. Активність публікації наукових співробітників є одним з основних факторів, який враховується при визначенні світових рейтингів вищих навчальних закладів. При цьому НБД є основними каналами подальшого застосування наукових результатів, як головної інформаційної та соціальної характеристики країни, університету, наукового колективу або окремого вченого.

Існуючі НБД, як правило, орієнтовані на пошук публікацій тільки у своїх сховищах. При цьому різні НБД використовують свої специфічні форми інтерфейсу, що визначає необхідність обов’язкової особистісної участі науковців у пошуку публікацій в різних базах. Ці обставини породжують протиріччя між необхідністю інтегральної оцінки публікаційної активності авторів і відсутністю інформаційних технологій, які дозволяють виконувати інформаційно-пошукові операції в різних НБД. Крім того існує задача багатоваріантного завдання атрибутів пошуку, у тому числі, різних варіантів написання прізвищ, оскільки деякі НБД використовують тільки англійські варіанти прізвищ. В різних НБД не використовуються моделі, методи та інструментальні методи визначення точності отриманої інформації, які засновані на аналізі прихованих змінних для виявлення зв'язків в наборі назв публікацій, що дозволяє ідентифікувати публікації конкретних авторів. Тому розробка інформаційно-пошукової системи по різних НБД є актуальною задачею, що спрямована на розв’язання вказаних протиріч у галузі інформаційних технологій.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалася відповідно до планів наукових досліджень ОНПУ, за участю автора як виконавця, за планами «Теорія і практика компетентнісного управління персоналом в організаційно-технічних і соціальних системах» (ДР № 0113U007624, 2012-2015) і «Методологія впровадження проектно-векторного управління інформаційними середовищами для моніторингу та управління науковими дослідженнями наукових груп» (ДР № 0115U000330, 2015).

Мета і задачі дослідження. Метою дослідження є розробка моделей, методів та інформаційної технології пошуку інформації для підвищення точності пошуку публікацій у наукометричних базах даних.

Для досягнення поставленої мети в дисертації поставлені і вирішені наступні завдання:

– аналіз структури метаданих та способів їх представлення у найбільш поширених і відомих наукометричних базах даних;

– аналіз структури пошукових систем з акцентом на інструменти завантаження веб-сторінок, та перетворення знайденої інформації у структурований формат даних;

– розробка методу отримання інформації із найбільш поширених наукометричних баз даних з можливістю розширення їх кількості;

– розробка моделі зберігання метаданих для подальшої їх обробки;

– розробка методів ранжування інформації з метою пошуку найбільш релевантних результатів;

– розробка інформаційно-пошукової системи із зручними інтерфейсами для завдання критеріїв пошуку та отримання результатів.

Об'єкт дослідження – процес пошуку інформації у наукометричних базах даних.

Предмет дослідження – моделі, методи та інструментальні засоби для створення інформаційно-пошукових систем.

Методи дослідження. Для досягнення мети і вирішення завдань, поставлених у дисертаційній роботі, використовується теоретичний аналіз способів доступу до інформації з наукометричних баз. Метод імітаційного моделювання використовується при побудові імовірнісних моделей з прихованими змінними (для ідентифікації публікацій конкретного автора).

Наукова новизна одержаних результатів.

– *вперше* розроблено комплексний метод пошуку та перетворення інформації із наукометричних баз даних у структурований формат даних, що дозволяє зменшити час на інтеграцію нових баз до системи;

– *вперше* розроблено методи ранжирування публікацій по їх назві з використанням імовірностних моделей латентно-семантичного аналізу і розміщення Діріхле, що дозволяє підвищити точність пошуку;

– *вперше* розроблена модель інформаційно-пошукової системи, особливістю якої є відсутність попереднього сканування повного вмісту наукометричних баз даних, що дозволяє скоротити використання обчислювальних ресурсів.

– *отримали подальший розвиток* інструментальні засоби інформаційної технологій пошуку та обробки інформації із динамічно створених на стороні користувача (клієнта) веб-сторінок, що значно розширює кількість придатних до обробки документів.

Практична значимість одержаних результатів. Розроблено програмний продукт «Science Metric Databases» (SMD), що реалізує інформаційну технологію пошуку даних у різних наукометричних базах даних. SMD, як складова інформаційно-пошукової системи метаданих публікацій науковців ВНЗ України, використана у НДР «Методологія впровадження проектно-векторного управління інформаційними середовищами для моніторингу та управління науковими дослідженнями наукових груп». Система дозволяє підвищити точність пошуку публікацій окремого автора у порівнянні зі схожими системами або пошуковими

машинами. Відкритість архітектури дозволяє розширювати можливості системи за рахунок інтеграції інших наукометричних баз даних. Отримані метадані публікацій використовуються програмним комплексом для моніторингу групової публікаційної активності лабораторій, кафедр і університетів. Система розгорнута на сервері кафедри Управління системами безпеки життєдіяльності ОНПУ (http://smd.opu.ua).

Особистий внесок здобувача. Всі результати наукових, теоретичних і практичних досліджень, які викладені у дисертації, отримані автором самостійно. У публікаціях, які опубліковані у співавторстві, використовувалися тільки ті положення та ідеї, які є результатом особистих досягнень аспіранта. Автором розроблено інформаційно-пошукову систему, що є результатами виконання даної роботи.

Апробація результатів дисертації. Основні результати роботи доповідалися та обговорювалися на конференціях і семінарах: міжнародна науково-практична конференція (МНПК) «Управління проектами: стан та перспективи» (Миколаїв - 2013); науково-методичний семінар «Шляхи реалізації кредитно-модульної системи» (Одеса - 2011, 2013, 2014, 2015); науковий семінар «Інформаційні технології в освіті, науці та виробництві» (Херсон - 2012, 2014); МНПК «Автоматизація: проблеми, ідеї, рішення» (Севастополь - 2014), МНПК «Молодь у світі сучасних технологій за тематикою: теоретико-методологічні та науково-практичні основи управління проектами підвищення конкурентоспроможності територій» (Херсон - 2014).

Публікації. Основні положення і результати дисертаційної роботи відображені в 18 публікаціях, з них 7 статей у наукових фахових виданнях України, які також індексуються у наукометричних базах даних, 8 статей у науково-технічних збірниках і 3 тези доповідей конференцій.

Структура дисертації. Дисертаційна робота складається з вступу, чотирьох розділів, висновків, списку використаних джерел з 83 найменувань на 10 сторінках, 2 додатків. Загальний обсяг основної частини становить 113 сторінок тексту, у тому числі містить 33 рисунка (1 з яких займає всю сторінку) і 13 таблиць (3 яких займають всю сторінку).

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтована актуальність досліджень, сформульовано проблему, мету і задачі дослідження, визначені об’єкт, предмет та наукова новизна дисертаційних досліджень. Відображена практична цінність досліджень. Наведено дані щодо апробації результатів дисертації та основні публікації.

У **першому розділі** розглядаються проблеми пошуку публікацій у наукометричних баз даних. Визначаються поняття наукометрії і наукометричних показників, наводиться характеристика найбільш відомих наукометричних баз даних і опис метаданих публікацій, що містяться в цих базах. Проводиться аналіз структури пошукових систем та доцільності розробки спеціалізованої системи для пошуку публікацій у наукометричних базах даних.

У сфері наукометричних вимірювань використовуються спеціалізовані інформаційні об'єкти, які називають наукометричними базами даних. Вони являють

собою засоби збереження та обробки наукометричних показників, а також, найчастіше, містять тіло публікаційних матеріалів (статей, журналів, книг), тобто є науковим репозиторієм. Найбільш авторитетні наукометричні бази даних світового рівня (Scopus, Web of Science) є реферативними базами даних, які включають в себе деяку кінцеву множину публікацій, а також засоби сервісу для задоволення інформаційних потреб кінцевих користувачів. При цьому, за деяким винятком, надання інформаційних послуг здійснюється на платній основі. Міжнародна практика наукометричних досліджень сьогодні найбільш часто базується на використанні двох баз даних: Web of Science і Scopus. Широко відомі також інші бази даних, які орієнтовані на інформаційне забезпечення наукових досліджень без формування даних наукометрії. Серед некомерційних наукометричних баз, в яких індексуються публікації українських вчених, можна назвати наступні: Copernicus, BASE, DOAJ, Driver, Science Index та інші.

У більшості випадків, наукометричні бази даних не містять в собі повного змісту наукової публікації, а тільки інформацію про неї і посилання на вихідний документ. Цю інформацію називають метаданими публікації. Метадані – це дані про дані або структуровані дані, що представляють собою характеристики описуваних сутностей для цілей їх ідентифікації, пошуку, оцінки, управління ними. Відомі способи стандартизації метаданих публікацій для полегшення можливої обробки їх автоматизованими засобами. Одним з них є використання спеціальних репозиторіїв, які призначені для документообігу певного типу (Eprints, Dspace, Digital Commons, OJS).

Для пошуку інформації, в основному в мережі Internet, створені пошукові системи. Вони складаються з пошукової машини і інтерфейсу користувача. Пошукова машина – це комплекс програм, призначених для пошуку інформації. Існує 4 типа пошукових систем:

– системи, що використовують пошукових роботів - складаються з трьох частин (рис. 1): краулер («робот», «бот»), індекс і аналізатор пошукових запитів. Краулер використовується для обходу мережі, завантаження документів і створення індексу їх вмісту. Частину краулера, що відповідає за завантаження та перетворення інформації у структурований формат даних називають павуком. Індекс – база даних контенту документів з посиланнями на вихідні документи. Аналізатор пошукових запитів – програма для розбору запиту і видачі результатів на основі індексу. Пошуковий робот в цьому механізмі постійно досліджує мережу, що робить інформацію більшою мірою актуальною. Більшість сучасних пошукових систем є системами даного типу;

– системи, керовані людиною (каталоги ресурсів) - ці пошукові системи заповнюються списком документів вручну; каталог містить метадані та короткий опис документа;

– гібридні системи - такі пошукові системи поєднують в собі функції систем, що використовують пошукових роботів і систем, керованих людиною;

– мета-системи - об'єднують і ранжирують результати відразу декількох по шукачів; ці пошукові системи корисні, коли у кожної пошукової системи є унікальний індекс.

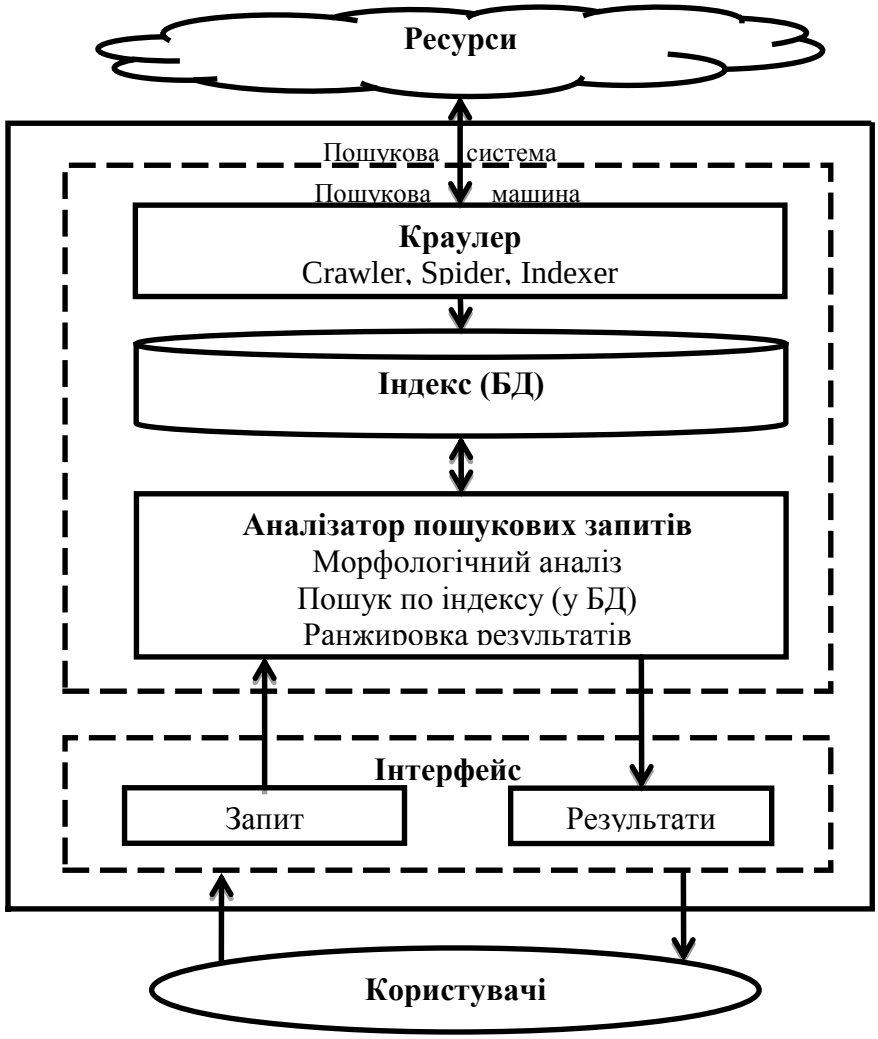

Рисунок 1 - Структура стандартної пошукової системи

На основі проведеного аналізу встановлено, що існуючі пошукові системи не працюють зі змістом наукометричних баз даних, а їх структура передбачає створення індексу всіх публікацій для подальшого використання його у процесі пошуку. Недоліками такого способу являються досить високі вимоги до обчислювальних ресурсів та необхідність вільного доступу до вмісту НБД. На підставі цього зроблено висновок про необхідність розробки спеціалізованої інформаційно-пошукової системи для пошуку публікацій у НБД, яка позбавлена вказаних недоліків.

У **другому розділі** розроблено модель пошуку і завантаження інформації із наукометричних баз даних, а також комплексний метод перетворення її у структурований формат даних. Зазначено проблеми, які виникали під час розробки моделі і методу та їх вирішення.

Зараз налічується значна кількість НБД, які розрізняються структурою і способом зберігання інформації. Проте існує один інтерфейс, який надають більшість НБД, хоча він орієнтований більше на користувачів, ніж на програмне забезпечення. Запропоновано використовувати веб-інтерфейс для доступу до вмісту (інколи в обмеженому вигляді) НДБ. Користувач, за допомогою веб-браузера, завантажує веб-сторінку певної НБД і, використовуючи пошук за заданими параметрами, отримує необхідну інформацію на сторінці.

Обробка контенту веб-сторінки породжує задачу аналізу та ідентифікації слабоструктурованих даних. Слабоструктуровані представлення даних відрізняються відсутністю строгих структур таблиць і відносин в моделях реляційних баз даних, тим не менш, ця форма даних містить теги та інші маркери для відділення семантичних елементів, а також для забезпечення ієрархічної структури записів і полів в наборах даних.

Вилучення структурованих даних з веб-сторінок зводиться до вирішення наступних завдань: пошуку та отримання цільових сторінок для отримання даних (проблема навігації); розпізнаванню ділянок, що містять потрібні дані (проблема розпізнавання даних); пошуку структури знайдених даних (проблема пошуку загальної структури даних); забезпечення однорідності даних (проблема зіставлення атрибутів вилучених даних); об'єднанню даних з різних джерел (проблема об'єднання даних).

Запропоновано вилучати інформацію, що орієнтована на сприймання користувачем, програмним способом. Таким чином, імітується робота користувача, який завантажив би тисячі веб-сторінок і зібрав інформацію у єдиному місці. На основі структури пошукової системи розроблено модель пошуку публікацій у наукометричних базах даних, показану на рис. 2.

Рисунок 2 - Модель пошуку та завантаження інформації

Для перетворення завантаженої інформації у структурований формат даних використано процес веб-скрапінгу, що фокусується на перетворенні неструктурованих даних в мережі (наприклад, у форматі HTML) в структурований формат даних, який може бути проаналізований і збережений. На відміну від моделі пошукової машини, сканується вузьке коло веб-сторінок, задане початковими умовами і витягується тільки корисна інформація. Алгоритм роботи починається з програми траулера, яка переходить по заданій адресі для певної бази даних, виділяє всі посилання, присутні на веб-сторінці і переходить по певним посиланнях, виходячи із заданих наперед умов. Слідуючи по знайдених посиланнях, траулер перенаправляє сторінки павуку для їх завантаження. Павук завантажує веб-сторінки і перетворює отриману інформацію у структурований формат даних. Для роботи з

багатьма базами даних, розроблено комплексний метод завантаження та обробки інформації, що показаний на рис. 3.

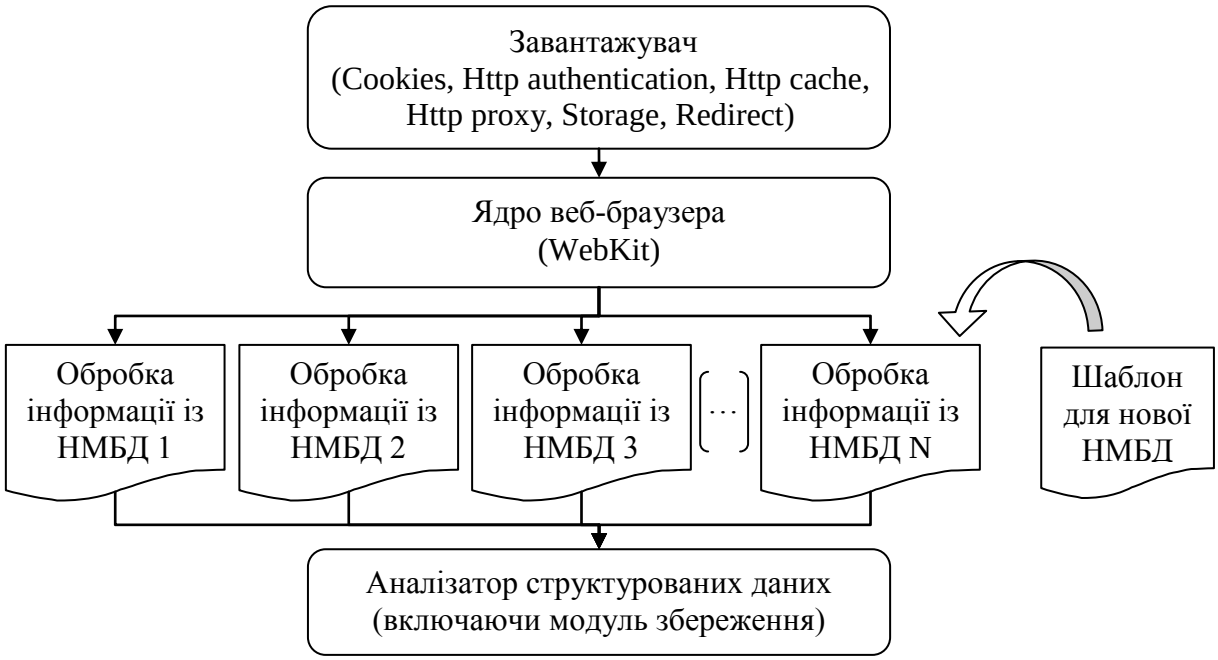

Рисунок 3 - Комплексний метод завантаження та обробки інформації із наукометричних баз даних

Особливістю даного методу являється можливість створення шаблону для обробки інформації із НБД, тому що основна частина кожного павука являється однаковою для всіх баз даних, змінюється лише частина, що відповідає за обробку структури конкретної веб-сторінки. Встановлено, що основна частина містить такі компоненти:

- завантажувач даних з мережі Інтернет;
- ядро веб-браузера для виконання програмного коду;
- аналізатор структурованих даних.

Під час розробки цього методу виникла проблема невідповідності контенту веб-сторінки завантаженої павуком і тієї ж веб-сторінки, але завантаженої за допомогою веб-браузера. Аналіз структури таких веб-сторінок виявив наявність програмного коду, що виконується браузером і модифікує результуючу веб-сторінку. Для вирішення цієї проблеми було удосконалено інструменти завантаження веб-сторінок за допомогою інтегрування ядра веб-браузера у модель завантаження інформації.

Після вилучення інформації у структурованому вигляді виконується подальша її обробка, що включає в себе фільтрацію результатів за деякими критеріями, групування публікацій по наукометричним базам, а також найбільш важливе і складне завдання – визначення однофамільців з метою підвищення точності результатів пошуку.

Таким чином запропонований інтерфейс доступу до НБД разом із комплексним методом завантаження та обробки інформації дозволяє отримати вміст НБД, а також швидко інтегрувати нові наукометричні бази даних.

У **третьому розділі** розроблено методи ранжування публікацій по їх назві на основі ключових слів, що підвищує точність пошуку. Розглянуто застосування імовірностних моделей з прихованими змінними, таких як латентно-семантичний аналіз і латентне розміщення Діріхле.

Аналіз результатів пошуку публікацій по автору виявив проблему однофамільців. Кількість знайдених публікацій може бути значно більшою ніж кількість існуючих публікацій конкретного автора, тому точність пошуку знижується. Для вирішення цієї проблеми застосовується тематичне моделювання для виявлення прихованих структур у текстових колекціях. Важливим є підклас орієнтованих імовірнісних тематичних моделей, які здійснюють м'яку кластеризацію і застосовуються для виявлення тематики текстів у колекціях документів. У термінах кластерного аналізу тема – це результат бі-кластеризації, тобто одночасної кластеризації і слів, і документів за їх семантичною близькістю. При м'якій кластеризації кожне слово і кожен документ належить до кількох тем одночасно з певними імовірностями. Таким чином, стислий семантичний опис слова або документа являє собою імовірнісний розподіл на множині тем. Процес знаходження цих розподілів називається тематичним моделюванням.

Для підвищення точності результатів пошуку публікацій розроблено два методи ранжування їх: на основі латентно-семантичного аналізу і на основі латентного розміщення Діріхле.

Латентно-семантичний аналіз (або латентно-семантичне індексування) – модель дворежимного факторного аналізу, яка базується на сингулярному розкладанні (SVD). Сингулярне розкладання представляє терміни та документи у вигляді векторів у просторі обраної розмірності, а скалярний добуток між точками простору – їх схожість. Латентно-семантичний аналіз починається з побудови матриці документів і термінів – індексованих слів. Індексовані слова – це слова, які включаються в двох або більше документах і мають смислове навантаження. Далі застосовується сингулярне розкладання цієї матриці на добуток трьох матриць:

$$A = U \cdot S \cdot V^t, \qquad (1)$$

де матриці U та V – ортогональні, а S – діагональна матриця, на діагоналі якої значення називаються сингулярними значеннями матриці A.

Таке розкладання володіє чудовою особливістю: якщо в матриці S залишити тільки *k* найбільших сингулярних значень, а в матрицях U і V – тільки відповідні цим значенням стовпці, то добуток одержаних матриць S, U і V буде найкращим наближенням початкової матриці A до матриці $\hat{A}$ рангу *k*:

$$\hat{A} \approx A = U \cdot S \cdot V^t, \qquad (2)$$

Основна ідея латентно-семантичного аналізу полягає в тому, що якщо в якості матриці A використати матрицю індексованих слів на документи, то матриця $\hat{A}$, що містить тільки *k* перших лінійно незалежних компонент A, відображає основну структуру різних залежностей, присутніх у вихідній матриці. Структура залежностей визначається ваговими функціями індексованих слів. Таким чином, кожне індексоване слово і документ представляються за допомогою векторів в загальному просторі розмірності *k*. Близькість між будь-якою комбінацією

індексованих слів та/або документів легко обчислюється за допомогою скалярного добутку векторів. Як правило, вибір k залежить від поставленого завдання і підбирається емпірично. Якщо вибране значення занадто велике, то метод втрачає свою потужність і наближається за характеристиками до стандартних векторних методів. Занадто маленьке значення k не дозволяє вловлювати відмінності між схожими термами або документами.

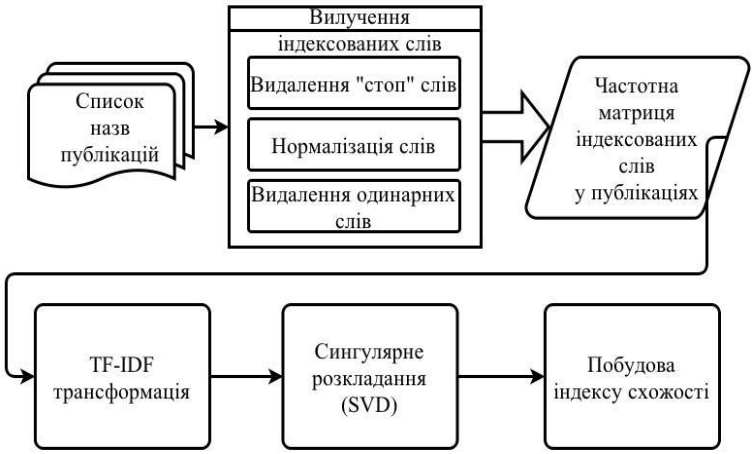

Рисунок 4 – Метод ранжирування на основі латентно-семантичного аналізу

Застосування методу латентно-семантичного аналізу для ранжирування публікацій показано на рис. 4. Спочатку треба аналізувати назви публікацій і виділити індексовані слова:

– видалення, так званих, "стоп" слів, тобто, які не мають смислового навантаження (прийменники, сполучники і т.д.);

– приведення слів до нормального вигляду або стемінг – процес знаходження основи слова (використано алгоритм Портера, який дозволяє швидко визначити основу слова);

– видалення слів, що зустрічаються тільки один раз.

На основі отриманих індексованих слів будується частотна матриця використання цих слів (рис. 5). Для підвищення якості аналізу, наступний етап – трансформація матриці за допомогою моделі TF-IDF (від англ. TF – term frequency, IDF – inverse document frequency) – статистична міра, яка використовується для оцінки важливості слова в контексті документа, що є частиною колекції документів. Вага певного слова пропорційна числу вживання цього слова в документі, і обернено пропорційна частоті вживання слова в інших документах колекції.

індексовані слова \ документи	1	2	3	4	5	6	7	8	9
1	1	0	1	0	0	0	0	0	0
2	0	0	1	0	0	0	1	0	0
3	1	0	1	0	0	0	1	0	0
4	0	1	0	1	0	0	0	0	0
5	0	1	0	0	0	0	1	0	0
6	0	0	0	1	0	1	0	1	0
7	0	1	0	1	0	1	0	0	0
8	1	0	1	0	0	0	0	0	0
9	1	0	1	0	1	0	1	0	1
10	1	0	1	0	0	0	0	0	0

Рисунок 5 – Матриця наявності індексованих слів в документах

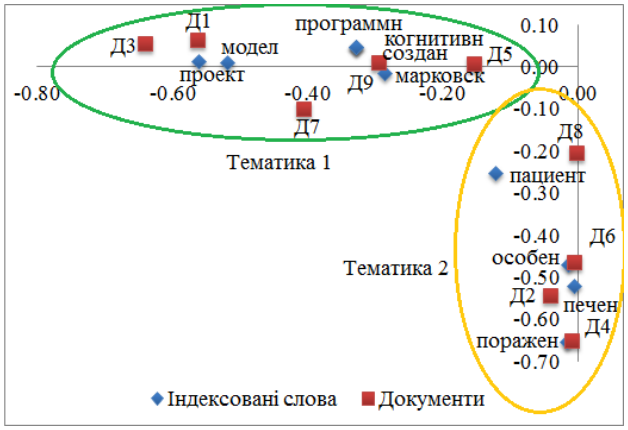

Рисунок 6 – Графічне представлення розподілу індексованих слів і документів у двовимірному просторі

Наступний крок, є основою латентно-семантичного аналізу – це сингулярне розкладання отриманої матриці і побудова індексу схожості, який обчислюється по відстані між індексованими словами і документами в k-вимірному просторі (рис. 6).

Другий метод ранжирування публікацій базується на латентному

розміщенні Діріхле (latent Dirichlet allocation, LDA). LDA передбачає, що кожне слово в документі породжене деякою прихованою темою. При цьому в явному вигляді моделюється розподіл слів у кожній темі, а також апріорний розподіл тем у документі.

Теми всіх слів у документі вважаються незалежними. LDA задає модель породження, як слів, так і документів, тому з'являється додаткова можливість оцінювати імовірності документів поза текстової колекції за допомогою алгоритму варіаційного виводу і семплірования по Гіббсу.

Застосування LDA дає модель, показану на рис. 7, де показано розподіл слів у двох темах. Виходячи з цього, можна судити, що документ 1 більше відноситься до першої теми, ніж до другої (30% + 20% + 10% проти 25%), а також отримати список найбільш підходящих до теми слів.

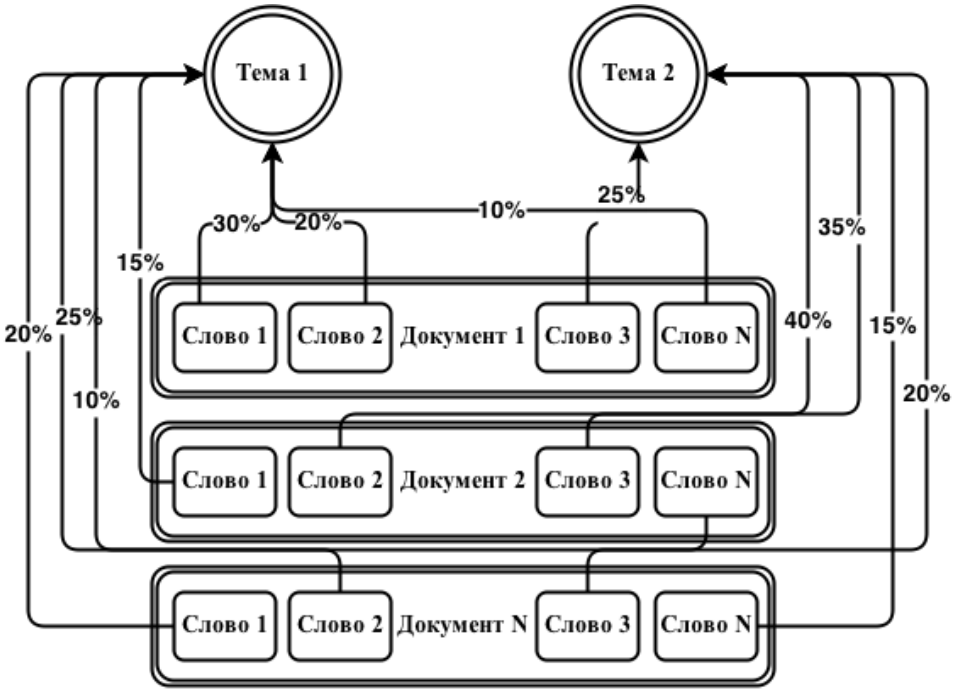

Рисунок 7 – Метод ранжирування на основі латентного розміщення Діріхле

Для ідентифікації параметрів моделі LDA по колекції документів можна застосувати семплірування по Гіббсу – алгоритм для генерації вибірки спільного розподілу множини випадкових величин. Він використовується для оцінки спільного розподілу і для обчислення інтегралів методом Монте-Карло.

Припустимо, що слід визначити *K* тем в наборі документів, тоді алгоритм семплірування по Гіббсу можна описати так:

1. Для кожного слова з кожного документа привласнити випадковим чином одну тему (t) з K можливих;
2. Для кожного слова з кожного документа обчислити:
   - p (t | d) – пропорція слів у документі d, які присвоєні темі t;
   - p (w | t) – пропорція слова w у всіх документах, присвоєного до теми t;
   - привласнити слову w нову тему t з імовірністю p (t | d) * p (w | t).
3. Повторити другий пункт кілька разів (кількість ітерацій також є вхідним параметром).

Метод LDA заснований на наступній імовірностній моделі:

$$p(d, w) = \sum_{t \in T} p(d) \cdot p(w \mid t) \cdot p(t \mid d), \quad (3)$$

де d – документ, t – тема, w – слово, T – множина тем, p (d) – апріорний розподіл на множині документів, p (w | t) – умовний розподіл слова w в темі t, p (t | d) – умовний розподіл теми t в документі d.

Розроблені методи ранжування дозволяють підвищити точність пошуку, вирішити проблему однофамільців а також видалити дублікати із результатів пошуку. Для практичного використання розроблених методів необхідно розробити інформаційну систему.

У **четвертому розділі** описано процес розробки інформаційно-пошукової системи для пошуку і обробки інформації із наукометричних баз даних.

Завданням даного програмного продукту є надати список публікацій здобувача, які індексуються в міжнародних наукометричних базах даних.

Вимогами до даного проекту є:

- витяг інформації з веб-сторінок;
- критерієм інформації є ПІБ автора;
- робота з найбільш поширеними наукометричними базами даних;
- обробка результатів з метою визначення нерелевантної інформації;
- надання інформації користувачеві.

Декомпозиція структури програмного проекту у вигляді набору компонент представлена на рис. 8.

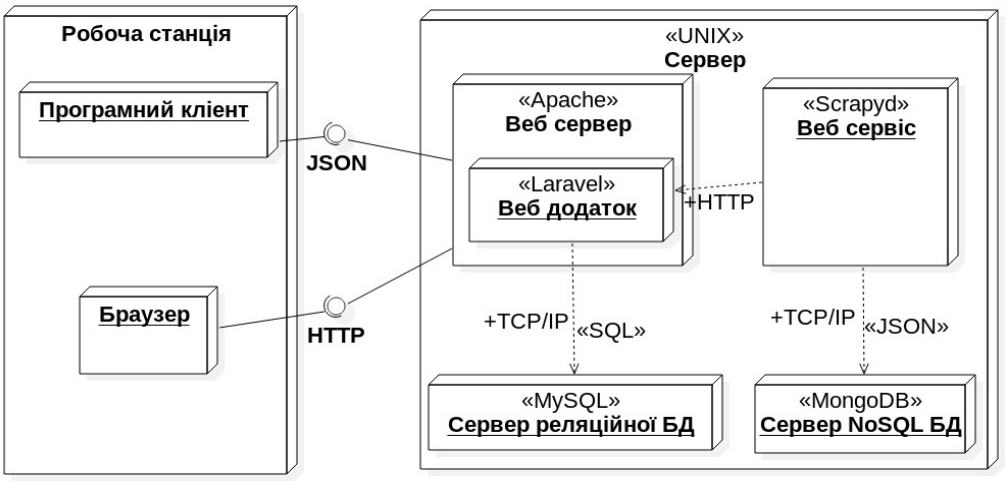

Рисунок 8 – Архітектура програмного проекту з вилучення публікацій

Проект являє собою програмний комплекс з декількох додатків які взаємодіють між собою. Основними компонентами системи є веб-додаток SMD та веб-сервіс Scrapyd. Додатковими компонентами є сервер реляційної БД MySQL і сервер NoSQL БД MongoDB. Веб додаток SMD являє собою графічний інтерфейс користувача, що дозволяє переходити до пошуку публікацій в певних базах (рис. 9).

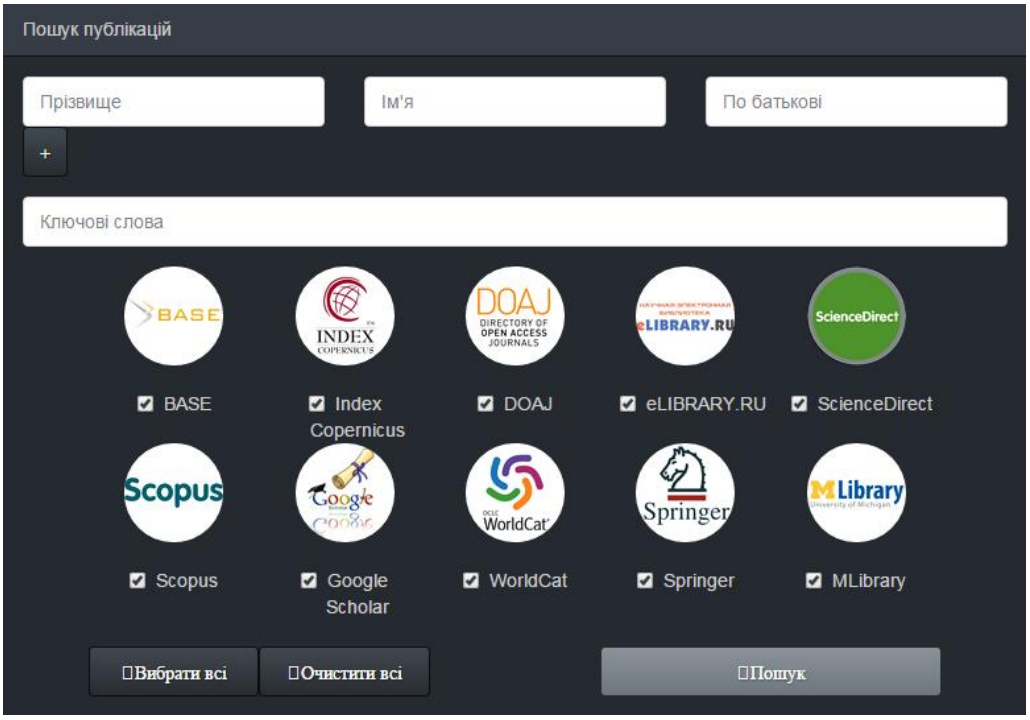

Рисунок 9 – Інтерфейс пошуку публікацій

Веб сервіс Scrapyd представляє сервіс по вилученню структурованих даних з НБД, а також управляє запуском відповідних програм-павуків окремої для кожної НБД. Таким чином, функціонал програмної системи розділений на окремі модулі – додатки, які працюють незалежно один від одного. Веб додаток SMD використовує сервіс Scrapyd під час для пошуку публікацій за запитом користувача. Ці програми спілкуються між собою по HTTP протоколу в JSON форматі.

Веб додаток SMD використовує реляційну базу даних (MySQL) в якості сховища даних, таких як інформація про користувачів, список підтримуваних НМБД, історія результатів пошуку публікацій та ін. Веб сервіс Scrapyd використовує документо-орієнтовану базу даних (NoSQL) для тимчасового зберігання результатів пошуку на зовнішньому диску, таким чином, не збільшуючи об'єм використання оперативної пам'яті при витяганні великої кількості публікацій. Доступ до баз даних надають окремі додатки – СУБД, з якими програми працюють по протоколу TCP/IP. Робота з додатком виконується за допомогою веб-браузера. Передбачено програмний доступ до інтерфейсу у форматі JSON. Основними варіантами використання програми є:

– реєстрація користувачів в системі – створення облікового запису користувача для можливості прив'язки знайдених публікацій до користувача;

– історія пошуку публікацій – навігація по списку пошукових запитів;

– перегляд і аналіз результатів пошуку (рис. 10);

– пошук публікацій – основний варіант використання; користувач запускає пошук за заданими параметрами з одночасним запуском сервісу пошуку (scrapyd), який керує цим процесом. Основні етапи пошуку публікацій це вилучення інформації, її аналіз (включаючи латентно-семантичний) і збереження результатів;

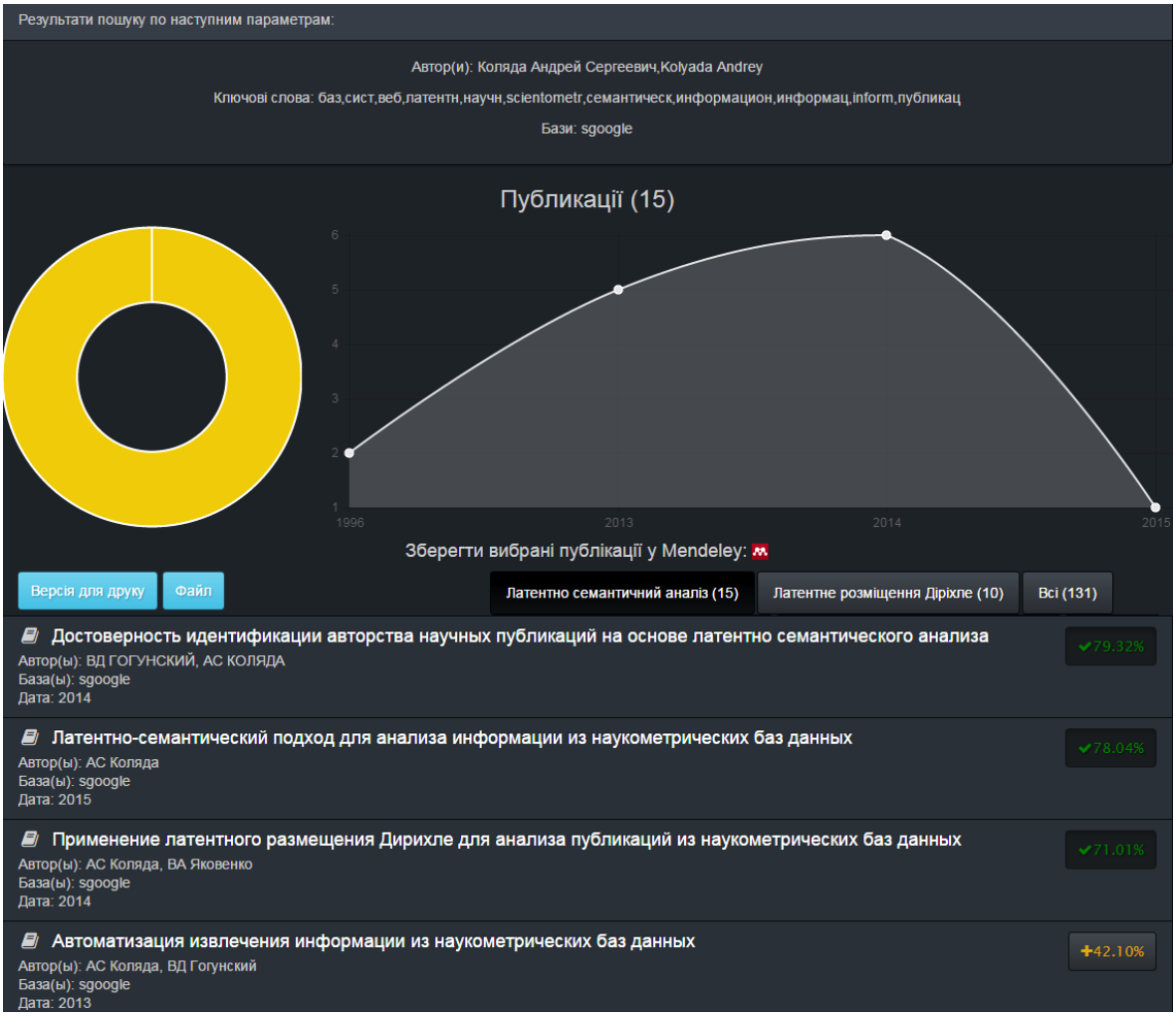

Рисунок 10 – Результати пошуку публікацій із застосуванням методу ранжирування на основі латентно-семантичного аналізу

– прив'язка публікацій до користувача і відображення статистики за знайденими публікаціями або публікаціями прив'язаних до користувача.

Сторінка графічного інтерфейсу результатів пошуку публікацій включає в себе інформацію про критерії пошуку, кількість знайдених всього публікацій та ідентифіковані публікації автора за допомогою методів ранжування, описаних у третьому розділі. Підтримується експорт публікацій у систему Mendeley (рис. 10). Передбачено також відображення нерелевантних результатів пошуку, оскільки публікації автора можуть відноситись до різних областей знань. Подальша робота з результатами можлива за допомогою їх друку або збереження у текстовий файл. Для підвищення точності пошуку публікацій автора підтримується автоматичне формування найвагоміших ключових слів із назв вибраних публікацій та збереження їх у профілі автора. Наступний пошук використає ці ключові слова для ранжирування результатів пошуку публікацій.

Із відомих аналогів даного програмного продукту можна зазначити програму Publish Or Perish, яка працює лише з двома джерелами публікацій – Google Scholar і Microsoft Academic Search. Тому порівняння результатів пошуку відображені з використанням одного джерела – Google Scholar.

Для оцінки результатів пошуку використано такі міри:

1. Точність – визначається як відношення числа релевантних документів, знайдених системою, до загального числа знайдених документів:

$$P = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}, \quad (4)$$

де D_{rel} – це множина релевантних документів у базі; D_{retr} – множина документів, знайдених системою.

2. Повнота – відношення числа знайдених релевантних документів, до загального числа релевантних документів базі:

$$R = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|}, \quad (5)$$

де D_{rel} – це множина релевантних документів у базі, а D_{retr} – множина документів, знайдених системою.

3. Міра Ван Різбергена, або F-міра – міра для спільної оцінки точності і повноти, яка визначається як зважене гармонійне середнє точності P і повноти R:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \beta = \frac{(1-\alpha)}{\alpha}, \alpha \in [0,1], \beta \in [0,\infty] \quad (6)$$

При $\alpha = 1/2$ або $\beta = 1$ F-міра надає однакову вагу точності і повноті і називається збалансованою мірою, формула для неї спрощується:

$$F_1 = \frac{2PR}{P + R} \quad (7)$$

Число релевантних публікацій у базі Google Scholar тестового автора є 17. У табл. 1 наведено оцінки результатів пошуку для порівняння двох систем.

Таблиця 1 – Порівняння результатів пошуку публікацій

Система		Знайдено релевантних публікацій/всього	Точність	Повнота	Міра Ван Різбергена
Publish Or Perish		15/155	0.096	0.882	**0.173**
SMD	LSA	12/15	0.800	0.706	**0.750**
	LDA	9/10	0.900	0.529	**0.666**

На основі цих оцінок можна зробити висновок щодо більшої ефективності розробленого програмного комплексу у порівняні з програмою Publish Or Perish. Значення збалансованої F-міри SMD у кілька разів перевищує показник Publish Or Perish.

Даний програмний продукт розроблено як один з інструментів інформаційного забезпечення моніторингу публікаційної активності науковців. Він надає метадані публікацій, які індексуються в міжнародних наукометричних базах даних. Основними особливостями програмного продукту є: можливість вилучення інформації з неструктурованих даних (веб-сторінок) і обробка цієї інформації з метою визначення нерелевантної інформації та фільтрації її.

ЗАГАЛЬНІ ВИСНОВКИ

В дисертації вирішена актуальна науково-прикладна задача теоретичного обґрунтування моделей і методів аналізу контенту веб-сторінок з імітацією роботи користувачів для автоматизованого вилучення метаданих наукових статей з наукометричних баз за допомогою розроблених програмних інструментів.

Отримані наступні результати.

1 *Внесок у теоретичні основи інформаційних технологій:*

1.1 На основі аналізу опублікованих робіт та існуючих програмних продуктів і інструментів наукометричних баз встановлено, що пошук публікацій в наукометричних базах даних, як правило, здійснюється тільки в межах окремих баз даних або репозитаріїв, що не дозволяє визначити інтегральну оцінку публікаційної активності науковців.

1.2 Виконана формалізація інформаційної технології для задач управління пошуком метаданих публікацій в наукометричних базах даних, що включає сучасну комп'ютерну систему накопичення, переробки і збереження інформації, що дозволяє розробити і впровадити Інтернет-технологію для побудови сервіс-орієнтованої системи інформаційного забезпечення кінцевих користувачів;.

1.3 Удосконалено метод Діріхле та модель латентно-семантичного аналізу, що містять ймовірнісні оцінки та інструментальні засоби класифікації і визначення достовірності інформації, що вилучається з контенту веб-сторінок, і засновані на аналізі прихованих змінних для виявлення зв'язків в наборі назв публікацій, що дозволяє достовірно ідентифікувати публікації конкретних авторів.

2 *Внесок в методи побудови інформаційно-пошукових систем:*

2.1 Запропонована концепція побудови інформаційно-пошукових систем і способів інформаційного забезпечення користувачів, яка базується на інформаційній технології вилучення та аналізу контенту веб-сторінок наукометричних баз даних, що дозволяє виконувати моніторинг інтегральної публікаційної активності, як окремих науковців, так і наукових колективів.

2.2 Розроблені програмні інструменти вилучення інформації з веб-сторінок, які конструюються динамічно на стороні користувача (клієнта), що дозволяє побудувати інформаційну технологію витягання контенту з елементами інтелектуальності в умовах невизначеності. Обґрунтовано і розроблено інформаційно-пошукову система автоматизації вилучення метаданих публікацій з поширених наукометричних баз даних, яка включає розроблені програмні інструменти вилучення та аналізу контенту веб-сторінок, що дозволяє виконати інтегральну оцінку публікаційної активності авторів наукових публікацій;

2.3 Розроблено програмний продукт, який реалізує інформаційну технологію пошуку публікацій науковців у найбільш відомих наукометричних базах даних; програмний продукт може бути корисним, як навчальним закладами, так і окремим науковцям, яким потрібно знати які їх публікації індексуються певними наукометричними базами даних.

3 *Створення передумов для подальших досліджень:*

3.1 Результати дисертаційних досліджень можуть бути основою для розвитку інформаційних технологій щодо забезпечення інформаційних потреб окремих

науковців зі створенням інформаційно-пошукових систем для більшого числа наукометричних баз даних.

3.2 Запропонована і розроблена інформаційна технологія, яка в роботі орієнтована на забезпечення особистих інформаційних потреб окремих науковців, може бути формалізована, як програмний інтерфейс (API), для включення в інші програмні комплекси для моніторингу публікаційної активності лабораторій, кафедр, університетів.

ОСНОВНІ ПУБЛІКАЦІЇ ПО ТЕМІ ДИСЕРТАЦІЇ

Статті у фахових виданнях України

1. **Коляда, А. С.** Автоматизація вилучення інформації із науко метричних баз даних [Текст] / А. С. Коляда, В. Д. Гогунський // Управління розвитком складних систем. – 2013. - № 16. – С. 96 – 99.

Видання включено до наукометричних баз даних (НБД): BASE; Index Copernicus.

Особистий внесок: виконано представлення і аналіз моделі даних у НБД.

2. **Коляда, А. С.** Вилучення інформації із слабоструктурованих веб-сторінок [Текст] / А. С. Коляда, В. Д. Гогунський // Східно-Європ. журнал передових технологій. - № 1/9 (67). – Харків : Технолог. центр, 2014 – С. 51 – 54.

[Видання включено до МНБ – Index Copernicus; Science Index; DOAJ; WorldCat; Ulrich's; DRIVER; BASE; Electronic Journals Library].

Особистий внесок:розроблено спосіб вилучення інформації із слабоструктурованих веб-сторінок на прикладі наукометричних баз даних

3. **Коляда, А. С.** Латентно-семантичний підхід для аналізу інформації із наукометрических баз даних [Текст] / А. С. Коляда // Управління розвитком складних систем. – 2014. – Вып. 17. – С. 90 – 94.

Видання включено до наукометричних баз даних (НБД): BASE, Copernicus.

4. **Коляда, А. С.** Достовірність ідентифікації авторства научних публікацій на основі латентно-семантичного аналізу [Текст] / А. С. Коляда, В. Д. Гогунський // Східно-Європ. журнал передових технологій. - № 3/2 (69). – Харків : Технолог. центр, 2014 – С. 36 – 40.

[Видання включено до МНБ – Index Copernicus; Science Index; DOAJ; WorldCat; Ulrich's; DRIVER; BASE; Electronic Journals Library].

Особистий внесок:розроблено модель визначення авторства наукових публікацій.

5. **Коляда, А. С.** Применение латентного размещение Дирихле для анализа публикаций из наукометрических баз данных [Текст] / А. С. Коляда, В. А. Яковенко, В. Д. Гогунський // Праці Одес. політехн. ун-ту. - 2014. – № 1 (43). – С. 186 – 191.

[Видання включено до МНБ –Science Index; Index Copernicus].

Особистий внесок: запропоновано використання моделі латентного розміщення Діріхле і проведено порівняльний аналіз із латентно-семантичним індексуванням.

6. Гогунський, В. Д. Наукометричні дані наукового видання «Управління розвитком складних систем» [Текст] / В. Д. Гогунський, **А. С. Коляда**, В. А. Яковенко // Управління розвитком складних систем. – 2014. - № 19. – С. 6 – 11.

Видання включено до наукометричних баз даних (НБД): BASE, Index Copernicus.

Особистий внесок: проаналізовано публікації наукового видання щодо подання його вмісту в наукометричних базах Index Copernicus і BASE.

7. Gogunsky, V. D. The development of the system concept of scientometric databases / V. D. Gogunsky, V. O. Iakovenko, **A. S. Kolyada** // Management of Development of Complex Systems. – 2014. - № 20. – pp. 143 – 147.

Видання включено до наукометричних баз даних (НБД): BASE, Index Copernicus.

Особистий внесок: проаналізовано принципи роботи наявних наукометричних баз даних.

Публікації апробаційного характеру

8. Гогунський, В. Д. Особливості цитування наукових публікацій у Інтернет-просторі / В. Д. Гогунський, В. О. Яковенко, А. С. Коляда // Наук.-метод. семінар: „Шляхи реалізації кредитно-модульної системи ”. – Вип. 10. — Одеса : Наука і техніка, 2015. — С. 28 – 33.

9. Коляда, А. С. Сучасні тенденції розвитку систем автоматизації технологічних процесів за наукометричними даними Scopus / А. С. Коляда, В. Д. Гогунський, В. А. Волобоєв // Наук.-метод. семінар: „Шляхи реалізації кредитно-модульної системи ”. – Вип. 10. — Одеса : Наука і техніка, 2015. — С. 46 – 52.

10. Гогунський, В. Д. Наукометричні бази: характеристика, можливості і завдання / В. Д. Гогунський, Г. О. Оборський, А. С. Коляда // Наук.-метод. семінар: „Шляхи реалізації кредитно-модульної системи”. – Вип. 8. — Одеса : Наука і техніка, 2014. — С. 3 — 12.

11. Коляда, А. С. Розробка програмного проекту для вилучення і обробки інформації із наукометричних баз даних / А. С. Коляда., В. Д. Гогунський // Інформ. технології в освіті, науці та виробництві : зб. – Вип. 2 (7). – Одеса : АО Бахва, 2014 – С. 191 – 195.

12. Гогунський, В. Д. Розробка наукометричних баз даних / В. Д. Гогунський, В. А. Яковенко, А. С. Коляда // Автоматизація: проблеми, ідеї, рішення: материали міжнар. науч. - техн. конф. Севастополь, 8-12 вересня 2014 г. / Севастоп. нац. техн. ун-т; науч. ред. В.Я. Копп – Севастополь : СевНТУ, 2014. – 184 с. – С. 111 — 113.

13. Гогунський, В. Д. Проектування системи моніторингу публікацій науковців в наукометричних базах даних / В. Д. Гогунський, А. С. Коляда, В. А. Яковенко // «Молодь у світі сучасних технологій» за тематикою» // Матеріали III Міжнар. наук.-практ. конф. (Херсон, 5-6 червня 2014 р.) / за заг. ред. Н.А. Соколова. Херсонський нац. техн. ун-т. – Херсон : ХНТУ, 2014. – С. 28 – 33.

14. Коляда, А. С. Розробка проекту інформаційно-аналітичної системи вилучення і обробки інформації із наукометричних баз даних / А. С. Коляда, А. А. Негрі, Е. В. Колеснікова // Управління проектами: стан та перспективи. IX Міжнар. наук.-практ. конф. — Миколаїв : НУК, 2013. — С. 348.

15. Коляда А. С. Запобігання інформаційної небезпеки / А. С. Коляда, В. Д. Гогунський // Наук.-метод. семінар: „Шляхи реалізації кредитно-модульної системи ”. – Вип. 7. — Одеса : Наука і техніка, 2013. — С. 54 — 57.

16. Коляда А. С. Розробка інтелектуальних систем навчання із застосуванням програмних агентів / А. С. Коляда, В. Д. Гогунський // Наук.-метод. семінар: «Шляхи реалізації кредитно-модульної системи …». – Вип. 5. — Одеса : ОНПУ, 2011. – С. 45 – 49.

17. Коляда, А. С. Ефективність використання адаптивних підходів при розробці програмного забезпечення / А. С. Коляда, С. Н. Ковешніков, В. Д. Гогунський //

Інформ. технології в освіті, науці та виробництві : зб. наук. праць. – Вип. 1. – Одеса : АО Бахва, 2012. - С. 29 – 33.

18. Коляда, А. С. Латентно-семантичний аналіз інформації із наукометричних баз / А. С. Коляда // Наук.-метод. семінар: „Шляхи реалізації кредитно-модульної системи ”. – Вип. 9. — Одеса : Наука і техніка, 2014. — С. 30 – 36.

АНОТАЦІЯ

Коляда А. С. Моделі і методи пошуку інформації у наукометричних базах даних. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Одеський національний політехнічний університет МОН України, Одеса, 2015.

Дисертація присвячена вирішенню проблеми створення інформаційної технології для вилучення метаданих наукових публікацій із наукометричних баз даних на основі веб-інтерфейсу.

Дано визначення наукометричної бази даних, приведено характеристику найпоширеніших із них, а також способи використання інформації з цих баз даних. Розроблено модель вилучення інформації із слабо структурованих веб-сторінок та модель автоматизації процесу вилучення із багатьох наукометричних баз даних. Також удосконалено спосіб вилучення інформації із динамічних веб-сторінок, які потребують виконання програмного коду на стороні користувача.

Проаналізовано процес тематичного моделювання, та застосовано латентно-семантичний аналіз і латентне розміщення Діріхле до списку назв вилучених публікацій з метою розподілу їх на близькі за змістом теми.

Розроблено програмну систему автоматизації вилучення метаданих публікацій з найбільш поширених наукометричних баз даних разом із графічним інтерфейсом користувача для управління пошуком публікацій, їх перегляду та аналізу.

Ключові слова: наукометрія, публікація, вилучення, модель, латентний, семантичний, Діріхле, слабоструктурований, веб-сторінка, краулер, павук.

АННОТАЦИЯ

Коляда А. С. Модели и методы поиска информации в наукометрических базах данных. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 – информационные технологии. – Одесский национальный политехнический университет МОН Украины, Одесса, 2015.

Диссертация посвящена решению проблемы создания информационной технологии для извлечения метаданных научных публикаций из наукометрических баз данных на основе веб-интерфейса.

Дано определение наукометрической базы данных, а также перечислены наукометрические показатели. Приведены характеристики наиболее

распространенных наукометрических баз данных. Показаны способы использования информации из этих баз данных на основе рейтингов университетов мира. Определено понятие метаданных публикации и приведены примеры их представления в наукометрических базах данных. Рассмотрена структура поисковых систем в сети Интернет, которые состоят из поисковой машины и интерфейса пользователя.

Разработаны методы извлечения информации из слабо структурированных веб-страниц и автоматизации этого процесса из многих наукометрических баз данных. Извлечение структурированных данных из веб-страниц сводится к решению следующих задач: поиск и получение целевых страниц с исходными данными (проблема навигации); распознавание участков, содержащих нужные данные (проблема распознавания данных); поиск структуры найденных данных (проблема поиска общей структуры данных); обеспечение однородности данных (проблема сопоставления атрибутов извлеченных данных); объединение данных из различных источников (проблема объединения данных). Для преобразования загруженной информации в структурированный формат данных использовано процесс веб-скрапинга, что фокусируется на преобразовании неструктурированных данных в сети (например, в формате HTML) в структурированный формат данных, который может быть проанализирован и сохранен. В отличие от модели поисковой машины, сканируется узкий круг веб-страниц, заданный начальными условиями и извлекается только полезная информация. Также усовершенствован способ извлечения информации из динамических веб-страниц, которые требуют выполнения программного кода на стороне пользователя.

Проанализирован процесс тематического моделирования и разработаны методы ранжирования публикаций по их названию на основе ключевых слов, что повышает точность поиска. Латентно-семантический анализ и латентное размещение Дирихле применены для решения проблемы определения публикаций конкретного автора, различая однофамильцев. Список названий извлеченных публикаций в данном случае является набором текстов – входной параметр вероятностных тематических моделей.

Разработана программная система автоматизации извлечения метаданных публикаций из самых распространенных наукометрических баз данных вместе с графическим интерфейсом пользователя для управления поиском публикаций, их просмотра и анализа. Предусмотрено предоставление программного интерфейса к функционалу данной программной системы с целью использования другими программами.

Из известных аналогов данного программного продукта можно отметить программу Publish Or Perish, которая, правда, работает только с двумя источниками публикаций – Google Scholar и Microsoft Academic Search. Для оценки результатов поиска публикаций в обеих системах использованы следующие меры: точность, полнота и мера Ван Ризбергена. Исходя из полученных оценок, определено, что эффективность поиска разработанной системы с применением ранжирования результатов сравнительно выше, чем у Publish or Perish. Данная система предназначена в первую очередь для мониторинга групповой публикационной

активности лабораторий, кафедр и институтов и внедрена в Киевском национальном университете строительства и архитектуры.

Ключевые слова: наукометрия, публикация, извлечение, модель, латентный, семантический, Дирихле, слабоструктурированный, веб-страница, краулер, паук.

ANNOTATION

Kolyada A. S. Models and methods of information search in scientometric databases. – The manuscript.

The dissertation for obtaining the scientific degree of Candidate of technical sciences in specialty 05.13.06 – Information technologies. – Odessa national polytechnic university MES of Ukraine, Odessa, 2015.

The thesis is devoted to the problem of creating information technology to extract metadata from scientometric publications database based on a web interface. The definition of scientometrics database is shown as well as the most common characteristic of them, including how to use these databases.

Developed the method of extracting information from poorly structured web pages and automation of the extraction process from many sciencesmetric databases. A method of information extraction from dynamic web pages that require code execution on the user side is also shown. Analyzed the process of topic modeling. And latent semantic analysis with latent Dirichlet allocation applied to the names of publications in order to place them close in content topics.

Software system to automate metadata extraction from publications of the most common scientometric databases with graphic user interface was developed to manage search publications viewing and analysis.

Keywords: scientometrics, extraction, publication, semistructured, model, latent, semantic, Dirichlet, webpage, crawler, spider.