МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНА МЕТАЛУРГІЙНА АКАДЕМІЯ УКРАЇНИ

КОВИЛІН ЄГОР РОМАНОВИЧ

УДК 004.822:004.055:519.878(043.5)

МОДЕЛЬ ГЕНЕРАЦІЇ ВІДПОВІДЕЙ В ПОШУКОВИХ СИСТЕМАХ НА ОСНОВІ НЕСТРУКТУРОВАНОЇ БАЗИ ЗНАНЬ

01.05.02 – математичне моделювання та обчислювальні методи

АВТОРЕФЕРАТ
дисертації на здобуття наукового ступеня
кандидата технічних наук

Дніпро – 2020

Дисертація є рукописом.

Роботу виконано на кафедрі комп'ютерних наук та інформаційних технологій Дніпровського національного університету імені Олеся Гончара (м. Дніпро) Міністерства освіти і науки України.

Науковий керівник: кандидат технічних наук, доцент
Волковський Олег Степанович,
доцент кафедри комп'ютерних наук та інформаційних технологій Дніпровського національного університету імені Олеся Гончара, м. Дніпро

Офіційні опоненти: доктор технічних наук, професор
Кіріченко Людмила Олегівна,
професор кафедри прикладної математики Харківського національного університету радіоелектроніки, м. Харків

доктор технічних наук, професор
Шинкаренко Віктор Іванович,
завідувач кафедри комп'ютерних інформаційних технологій Дніпровського національного університету залізничного транспорту ім. академіка В. Лазаряна, м. Дніпро

Захист відбудеться «___»_____ _____ р. об __ годині на засіданні спеціалізованої вченої ради Д 08.084.01 у Національній металургійній академії України за адресою: 49600, м. Дніпро, пр. Гагаріна 4, НМетАУ, конференц-зал засідань.

З дисертацією можна ознайомитись у бібліотеці Національної металургійної академії України за адресою: 49600, м. Дніпро, пр. Гагаріна, 4, НМетАУ.

Автореферат розіслано « ___ » _____ 2020 р.

Вчений секретар
спеціалізованої вченої ради Тетяна СЕЛІВЬОРСТОВА

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність роботи. Задача пошуку інформації є однією з ключових в області комп'ютерних наук. Якщо в функціоналі системи присутня необхідність працювати безпосередньо з користувацьким запитом, який часто складається з декількох неформальних критеріїв і вимагає певного семантичного аналізу, то обробка отриманих результатів повністю лягає на плечі користувача. Прикладом такого підходу є модель, яка використовується у багатьох популярних web-пошукових системах: відповіддю на отриманий запит є множина документів, що припускає подальший самостійний аналіз користувачем кожного документа для пошуку відповіді на своє питання. Головним мінусом такого підходу є відсутність глибокого семантичного розуміння вмісту документа моделлю, у результаті чого в отриманому масиві документів міститься велика кількість не пов'язаної із запитом користувача інформації, а також множини повторень однакової інформації, поданої в різних інтерпретаціях, через що процес пошуку значно ускладняється.

Вирішенням цієї задачі є використання процесу автоматичної генерації знань, який дозволить відразу отримати релевантні до запиту користувача знання на основі усіх знань, що містяться у системі, та вдосконалити таким чином процес пошуку даних для людини-оператора. Головною складністю впровадження генерації знань у пошукові моделі є необхідність побудови семантичної моделі документів, доступної для обробки комп'ютером. Окремо слід відзначити ускладнення, пов'язане із властивостями флексії та типології порядку слів мови документа, через що моделі, створені наприклад для англійської мови, не є застосовними для документів на слов'янській мові. Через це, головним підходом до створення семантичних моделей документів на слов'янських мовах є ручна побудова семантичної структури тексту, яка може бути представлена у вигляді онтологічної розмітки або семантичних словників, і вимагає використання великого обсягу ручної праці для їх складання і пошуку фахівців у прикладній лінгвістиці, що значно ускладнює використання автоматичної генерації знань у пошукових системах та обмежує застосування моделі для довільної колекції семантично неструктурованих текстів.

Питання розробки схожих моделей представлення текстів для їх подальшої комп'ютерної обробки подані у багатьох роботах, які стосуються галузей штучного інтелекту, математичного моделювання та обробки природної мови. Серед вітчизняних вчених слід відзначити наукові розробки Н.Н. Леонтьевої, М.В. Мозгового, В.А. Тузова, І.О. Мельчука, Ю.Д. Апресяна, спрямовані на математичне моделювання семантичних властивостей тексту на основі онтологій та семантичних словників. Зарубіжні англомовні моделі представлення текстів відштовхуються від жорсткого порядку слів у мові, як наприклад в роботах N. Chomsky. В області вилучення знань із текстів панують онтологічні математичні моделі, описані в роботах B. Yildiz, S. Miksch, R. Navigli, S. Dill, L.K. McDowell, D. Faure.

Проведений аналіз наявної у відкритому доступі науково-технічної літератури і документації показав, що існуючі моделі представлення слов'яномовних текстів спрямовані саме на опис структурованих знань, і не дозволяють повністю

автоматизувати процес опису семантичних властивостей та адаптивного додавання неструктурованих знань до пошукової системи. Усі вони мають на увазі залучення значної кількості лінгвістичних знань і використання попередньої семантичної розмітки, створеної лінгвістом-експертом вручну. З іншого боку, зарубіжні моделі орієнтовані в першу чергу на обробку англомовних текстів і не можуть бути застосовані для представлення слов'яномовних текстів.

Таким чином, недосконаліть процесу пошуку інформації і відсутність цілком автоматичних моделей представлення слов'яномовних текстів дозволяє зробити висновок про те, що на сьогоднішній день задача розробки моделі генерації відповідей на основі неструктурованої бази знань в пошукових системах є актуальним завданням.

Зв'язок роботи з науковими програмами. Робота виконана відповідно до закону України «Про пріоритетні напрями розвитку науки і техніки» (Відомості Верховної Ради України (ВВР), 2001, № 48, ст.253), і стосується напряму «інформаційні та комунікаційні технології» (стаття 3). Обраний напрямок досліджень пов'язаний із виконанням дослідних робіт кафедри комп'ютерних наук та інформаційних технологій Дніпровського національного університету імені Олеся Гончара «Методи та інформаційні технології цифрової обробки багатоканальних даних» (реєстраційний номер 0116U001297).

Мета та задачі дослідження. Метою дисертації є розробка моделі обробки семантично-неструктурованих документів для генерації відповідей в пошукових системах.

Для досягнення мети були поставлені та вирішені наступні задачі дисертаційного дослідження:

1) визначити концепції моделі вилучення інформації та архітектури системи, для чого провести аналіз і порівняльну характеристику фундаментальних моделей уявлення текстових знань, з метою обґрунтованого вибору оптимальної для вирішуваної задачі;
2) розробити математичну модель представлення семантичних властивостей текстів, спроможну працювати із достатньо формалізованим типом тексту і автоматично формувати програмну семантичну модель як окремого документа, так і всього корпусу знань в цілому;
3) забезпечити розроблену модель можливістю використовувати семантично нерозмічений заздалегідь корпус текстів. Модель повинна функціонувати без використання лінгвістичних знань про мову або семантичні властивості текстів;
4) розробити математичну модель валідації текстів-кандидатів до включення у неструктуровану базу знань за ступенем їх семантичної зв'язності, яка захистить неструктуровану базу знань від наповнення помилковими даними;
5) на основі семантичної моделі та моделі валідації, розробити математичну модель генерації текстів, її алгоритмічне та програмне забезпечення;
6) розробити і застосувати систему оцінок адекватності створеної моделі генерації відповіді на основі неструктурованої бази знань.

Об'єкт дослідження. Процес автоматичної генерації відповідей з неструктурованої текстової бази знань.

Предмет дослідження. Модель генерації відповідей з неструктурованої бази знань на основі автоматичного вилучення семантичних характеристик тексту і попередньої класифікації даних.

Методи дослідження. Для вирішення поставлених задач були використані методи частотного аналізу текстів, факторного латентно-семантичного аналізу, кластерного аналізу, індукційного і дедуктивного аналізу, вимірювання, експерименту, гіпотези, припущення, комп'ютерного моделювання отриманих результатів. Застосовано засоби теорії множин, штучного інтелекту і проекційного методу.

При розробці програмної реалізації побудованих моделей були застосовані: технологія об'єктно-орієнтованого програмування, реляційна модель зберігання даних і засоби CASE-проектування архітектури програмного додатку. Для дослідження адекватності розроблених моделей використано статистичні методи і метод бальних оцінок.

Наукова новизна отриманих результатів. В роботі **вперше** отримано такі результати:

1) розроблено семантичну модель текстових даних, яка на відміну від існуючих аналогів, дозволяє отримувати кількісні показники семантичних властивостей і сенсові зв'язки між компонентами тексту без необхідності будь-якої попередньої семантичної розмітки, впровадження словників або залучення лінгвістичних знань;
2) створено модель автоматичної класифікації знань за ступенем їх семантичної зв'язності, що використовує числові дані, отримані із розробленої семантичної моделі тексту, яка дозволила збільшити надійність використання моделі генерації відповідей шляхом верифікації первинної інформації;
3) побудовано модель автоматичної генерації відповідей у пошуковій системі із неструктурованої бази текстових знань на основі створених моделей, яка дозволила автоматизувати роботу користувача із пошуковими системами;
4) розроблено систему оцінок адекватності створеної моделі генерації відповіді на основі неструктурованої бази знань;

Отримали подальший розвиток:

5) семантичні моделі текстових даних для флективно багатих мов із вільним порядком слів: розроблена семантична модель текстових даних дозволяє уникнути процесу ручного опису семантичної структури документа;
6) методи організації пошуку інформації: створені моделі дозволяють генерувати релевантні до запиту користувача знання на основі неструктурованої бази знань, чим спрощують роботу користувача із пошуковими системами.

Практичне значення отриманих результатів. Розроблено математичну модель генерації відповідей в пошукових системах на основі неструктурованої бази

знань та побудовано на її основі комп'ютерну систему, яка окрім організації зручного пошукового середовища, утворює універсальний програмний фреймворк з набором інструментів для проведення автоматичної обробки текстів на семантичному рівні, доступним для будь-якого користувача у вигляді окремої бібліотеки.

Формат реалізації додатку пояснюється актуальністю задачі через відсутність альтернативних API із відкритим доступом. Задачі дисертаційної роботи є важливими і нагальними питаннями галузі цифрового моделювання текстів, вирішення яких дозволяє:

1. гнучко створювати і обробляти тематичні повнотекстові бази знань без попередньої семантичної розмітки;
2. будувати програмну модель текстових знань формалізованої стильової спрямованості із кількісними характеристиками семантичних властивостей, на основі яких можливо вирішувати інші завдання автоматичної обробки текстів без необхідності залучати будь-які лінгвістичні знання.

Особистий внесок здобувача. Основні результати дисертаційної роботи опубліковано в статтях:

[1, 2] – аналіз існуючих розробок і моделей автоматичного уявлення текстових знань та оцінка можливості їх застосування для вирішення задачі побудови моделі генерації відповідей із неструктурованої бази знань; [3, 4] – розробка моделі автоматичної класифікації формалізованих текстових знань на основі моделі автоматичного уявлення семантичних характеристик тексту із неструктурованої бази текстових знань; [5] – побудова моделі і комп'ютерної системи інтелектуального семантичного пошуку з використанням генерації текстів; [6, 8] – побудова моделі автоматичного уявлення семантичних характеристик тексту із неструктурованої бази текстових знань; [7] – розробка моделі автоматичної оцінки адекватності комп'ютерних систем «запит-відповідь» з використанням генерації текстів та оцінка адекватності створеної пошукової моделі.

Апробація результатів дисертації. Основний зміст та висновки дисертаційної роботи викладені та обговорені на 6 міжнародних та всеукраїнських конференціях: міжнародній науково-технічній конференції «Інформаційні технології в металургії та машинобудуванні-2017» (м. Дніпро), міжнародній науково-практичній конференції «Інформаційні технології та комп'ютерне моделювання-2017» (м. Івано-Франківськ), міжнародній науково-технічній конференції «Інформаційні технології в металургії та машинобудуванні-2018» (м. Дніпро), міжнародній науково-технічній конференції «IEEE Second International Conference on Data Stream Mining-2018» (м. Львів, **НМБ Scopus**), XIX міжнародній науково-технічній конференції з математичного моделювання, присвяченої 250-річчю з дня народження Жозефа Фур'є (м. Херсон), всеукраїнській науково-методичній конференції «Проблеми математичного моделювання-2018» (м. Кам'янське).

Розроблений програмний додаток впроваджено у міський комунальний заклад культури «Централізована система бібліотек для дітей» м. Дніпро як пошуковий інструмент обробки електронних текстів, у ТОВ «Сітал Україна» як засіб

автоматичної генерації текстових інструкцій та у АТ «ДніпроАзот» як інструмент покращення процесів пошуку в системах електронного документообігу.

Публікації. Результати дисертаційної роботи опубліковані в 14 наукових працях, в тому числі 8 статей у журналах, рекомендованих МОН України для публікації результатів дисертацій, та закордонних виданнях: «Математичне моделювання» – 1 (НМБ Index Copernicus), «Системні технології. Регіональний збірник міжвузівських наукових праць» – 5 (НМБ Index Copernicus), «Вісник Херсонського національного технічного університету» (НМБ Google Scholar) – 1, «Modern engineering and innovative technologies» (Німеччина, НМБ Index Copernicus) – 1; у тезах доповідей та трудах міжнародних та всеукраїнських конференцій – 6.

Структура та обсяг дисертації. Дисертаційна робота викладена на 233 сторінках тексту, складається зі вступу, 3 розділів, загальних висновків, списку використаних джерел із 107 найменувань та 19 додатків. Обсяг основного тексту дисертації складає 146 сторінок тексту. Робота ілюстрована 16 таблицями та 46 рисунками.

ОСНОВНИЙ ЗМІСТ ДИСЕРТАЦІЇ

У **вступі** обґрунтовано актуальність обраної теми, її зв’язок з науковими програмами, визначено мету та задачі дослідження. Визначено та розкрито предмет, об’єкт і межі дослідження, висвітлені методи, які використані в процесі роботи. Розкрита наукова новизна та практичне значення одержаних результатів.

У першому розділі проаналізовано існуючі математичні моделі текстів та існуючі аналоги моделі генерації текстів на основі неструктурованої бази знань у пошуковій системі. Здійснено і обґрунтовано вибір базової концепції моделі та зазначені основні складнощі та наукові задачі, які пов’язані із її використанням у пошукових системах. Сформульована модель генерації відповідей із неструктурованої бази знань та основні кроки її роботи. Введено поняття «текстовий автомат» та визначені основні вимоги до функціонування розробленої пошукової моделі.

Розглянуті існуючі моделі текстів, що є кандидатами для застосування у основі моделі генерації текстів, а саме: модель граматичного опису, сформульована американським лінгвістом А.Н. Хомським, яка отримала назву генеративна граматика, альтернативна модель мови, що була сформульована групою вчених (І.А. Мельчук, Ю.Д. Апресян, А.К. Жовківський) в СРСР і отримала назву «Теорія Сенс ↔ Текст», модель м’якого автоматичного розуміння тексту, розроблена під керівництвом проф. Н.М. Леонтієвої (СРСР). Розкриті головні концепції кожної із моделей текстів – виділення синтаксичних структур у генеративній граматиці, робота із тлумачно-комбінаторним словником у теорії «Сенс ↔ Текст» та ситуативна зміна розуміння семантики у моделі м’якого розуміння тексту. Розглянуті основні переваги та недоліки обраних теорій з точки зору їх застосування до задачі автоматичної генерації відповідей у пошуковій системі.

Обґрунтовано вибір базової моделі обробки неструктурованих знань – моделі м'якого автоматичного розуміння, оскільки до плюсів такого підходу відносяться початкова орієнтованість на синтез, висока роль семантики в моделі, незалежність від синтаксичної структури тексту, вирішення проблеми флексії мови тексту та орієнтованість на спрощену базу знань. Визначені пріоритетні наукові питання, що поставлені у дисертації, пов'язані з організацією навчання моделі, подоланням обмежень семантичного словника і гнучким налаштуванням роботи моделі при підключенні нових предметних областей текстових даних.

Розкрите питання прикладної реалізації обраної базової алгоритмічної моделі мови. Сформульовані основні задачі цього етапу – створити складну багаторівневу систему аналізу даних, що містить у собі велику кількість допоміжних утиліт, наповнити модель знаннями, організувати автоматизований процес навчання і адаптації моделі, здійснити оцінки отриманих результатів, знизити зв'язність компонентів моделі, чітко сформулювати критерії до архітектури програмного додатку. Розкрита базова концепція текстового автомату як засобу мінімізації інформаційних втрат при здійсненні генерації відповідей у пошукових системах та перешкоджанні виходу системи з ладу – як з технічної, так і з логічної точки зору. Проаналізована доцільність використання текстового автомату у рамках моделі м'якого автоматичного розуміння – модулі, що входять до його складу, формують рівні процесу обробки текстових даних за своїм безпосереднім функціоналом, що створює пряму паралель між ієрархічним процесом роботи текстового автомату і рівнями розуміння знань моделі. Сформульована основна гіпотеза дисертаційної роботи, яка покладена у фундамент моделі – ідея про підходи до формування семантичних міток текстового знання моделі. Головним структурним компонентом генерації текстів на основі запиту була обрана семантична мережа, яка вказує на взаємозв'язок семантичних блоків, що складаються з залежних диференційованих вершин термінів – сенсових значень для конкретного тексту. Така модель дозволяє автоматично створювати базу знань, в якій містяться семантичні властивості кожного терміну без залучення лінгвістичних даних. Запропоновані основні кроки головного алгоритму роботи моделі. Визначені критерії до функціонування моделі генерації відповідей на основі неструктурованої бази знань в пошукових системах.

Проведено аналіз існуючих розробок і основних моделей, які можуть бути покладені в основу моделі і дозволять уникнути дублювання вже реалізованого функціоналу. Аналіз показав, що серед сучасних програмних засобів моделювання текстових знань напрямок генерації відповідей у пошукових системах є не дуже розвинутим, проте були знайдені системи, що підпадають під задачу дисертації. Їх дослідження показало, що модель генерації відповідей є актуальним та інноваційним засобом для генерації текстових знань у пошукових системах.

У другому розділі розкриті основні властивості семантичних мереж, критерії та алгоритм побудови семантичної моделі неструктурованого тексту, наведені результати автоматичної побудови семантичної моделі документа та проведена перевірка адекватності розробленого підходу.

Сформульовані головні критерії до створеної семантичної моделі, а саме: відсутність необхідності додавання семантичного словника знань, попередньої семантичної розмітки тексту, або будь-яких інших семантичних знань, що вимагають впровадження структури бази знань; отримана модель має спиратися саме на семантичні властивості текстових знань і має бути орієнтована на генерацію відповідей та роботу із запитом користувача; отримана модель має містити кількісні характеристики семантичних властивостей текстових знань для подальшого інтелектуального автоматичного аналізу. Розкрито алгоритм роботи та застосування латентно-семантичного аналізу, як базового підходу для побудови семантичної моделі текстових знань.

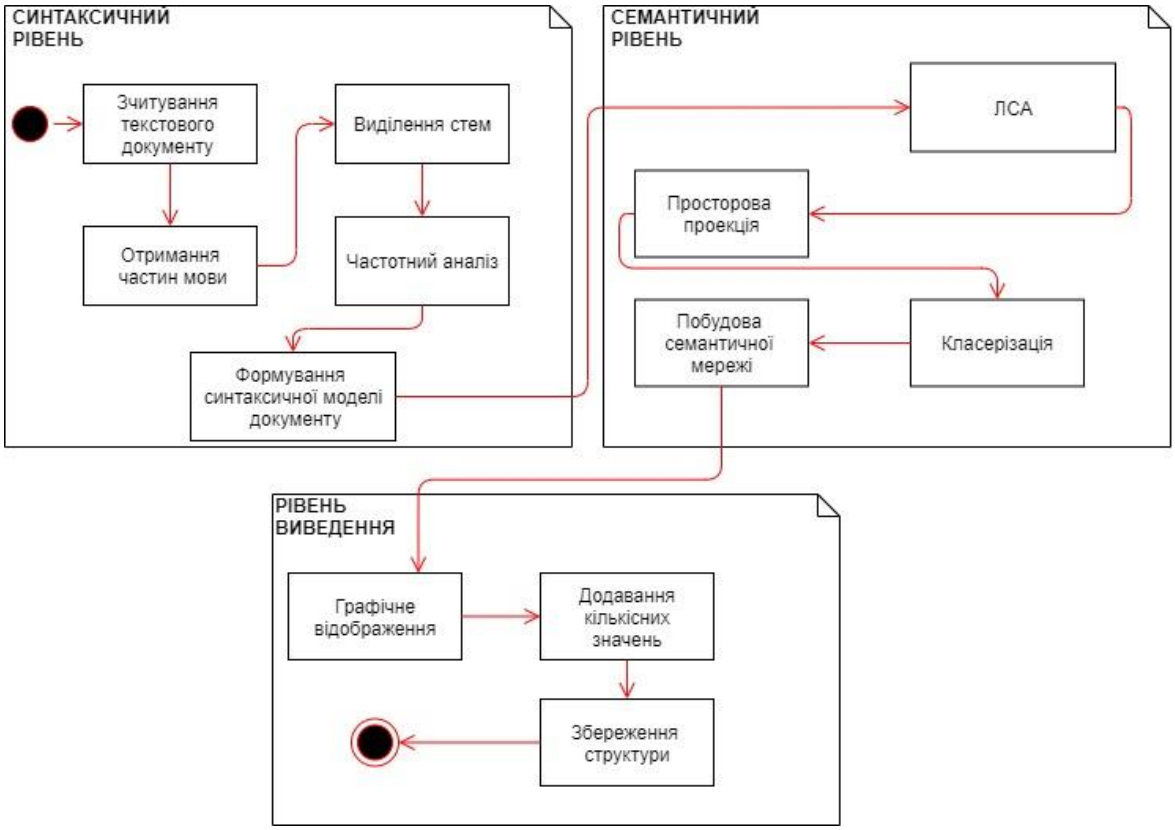

Рис. 1. Схематичне зображення процесу побудови семантичної моделі тексту

Загальний процес побудови моделі (рис. 1) має лінійну структуру, яка поділена на три рівня обробки документу:

– синтаксичний рівень, який об’єднує у собі деякі типові задачі обробки тексту, такі як виділення слів і речень із вхідного тексту, визначення стем, зважування елементів тексту, автоматичне визначення частини мови, тощо;

– семантичний рівень, який реалізує у собі компонент латентно-семантичного аналізу і додаткові засоби інтелектуальної обробки даних, завдяки яким відбувається формування фінальної семантичної моделі вхідного тексту;

– рівень виведення, який відповідає за відображення і збереження отриманої семантичної моделі у базу знань.

Першим завданням синтаксичного рівня, що виконується над вхідним текстом у форматі «plain text» (лише текст, без зображень, формул або таблиць, у якому відсутня будь-яка додаткова семантична розмітка), стає синтаксичний аналіз, який ставить перед собою за мету виділення речень та слів з тексту. Як тільки слова і речення переводяться у програмні сутності, наступною задачею стає очистка тексту від стоп-слів, що співвідносяться із неінформативними частинами мови. Оскільки застосування словника суперечить головній меті даної роботи, то було прийнято рішення про використання системи оцінки імовірності належності слова до тієї чи іншої частини мови на основі наївного байєсівського класифікатору. Для цього, в системі задається кінцевий набір пар $P<W,POS>$ де W – деяке слово, POS – його частина мови, набір стоп слів S, де $s \in S$ – має частину мови союз, частка, вигук або займенник, та набір значущих (informative) частин мови POS_I. В цьому випадку, термін T потрапляє в результуючий (resulted) набір T_R, якщо виконується умова:

$$T \in T_R \Leftrightarrow T \notin S \wedge B(last2(T), last3(T), P<W,POS>) \in POS_I, \quad (1)$$

де $B(...)$ – функція байєсівського класифікатора, яка повертає клас – частину мови, $last2(T)$– функція, яка повертає ознаку двох останніх символів слова, $last3(T)$ – функція, яка повертає ознаку трьох останніх символів слова.

Над отриманою множиною T_R виконується операція стемінгу слів – приведення однокореневих слів до деякої загальної форми – стеми. Для цього було прийнято рішення про розробку і використання підходу, який базується на знаходженні дистанції Левенштейна. Дистанцію Левенштейна можна сформулювати наступним чином: нехай є пара слів $S_1, S_2 \in T_R$, довжиною M и N відповідно. Тоді відстань Левенштейна $d(S_1, S_2)$ можна вирахувати за такою рекурентною формулою $D(M,N)$:

$$D(M,N) = \begin{cases} 0, i=0, j=0, \\ i, j=0, i>0, \\ j, i=0\, j>0, \\ \min\begin{cases} D(i,j-1)+1, \\ D(i-1,j)+1, \\ D(i-1,j)+m(S_1[i],S_2[j]) \end{cases}, i>0, j>0 \end{cases}, \quad (2)$$

де $m(a,b)$ – дорівнює нулю якщо $a = b$ і одиниці в іншому випадку, $\min\{a,b,c\}$ повертає найменший з аргументів.

Тут крок за i символізує видалення з першого рядка, по j – вставку в перший рядок, а крок за обома індексами символізує заміну символу або відсутність змін. На основі отриманої дистанції для кожної пари слів S_1, S_2 відбувається заміна одного рядка на інший до тих пір, поки істинно:

$$S_1 = S_2 \Leftrightarrow d(S_1, S_2) < len_{common}(S_1, S_2), \quad (3)$$

де $len_{common}(S_1, S_2)$ – довжина найбільшої загальної частини пари стем.

Відносна похибка цього підходу склала 2% і була вирахувана експериментально за формулою:

$$\delta = \left(1 - \frac{\sum N_S}{\sum N}\right) * 100\%, \quad (4)$$

де $\sum N$ – загальна кількість розглянутих пар стема до стеми-кандидату, $\sum N_S$ – кількість правильно встановлених розглянутих пар стема до стеми-кандидату.

Над нормалізованим таким чином текстом виконується зважування стем, у результаті чого кожній стемі відповідає кількість її повторень у тексті і зважування речень, де під вагою речення мається на увазі сумарна вага усіх стем речення.

Для організації інтелектуального розуміння семантики тексту у моделі виконується ряд обробок статистичних даних, отриманих на рівні синтаксичного аналізу, сукупність яких утворює семантичний рівень моделі. Розмічений текст D, отриманий на синтаксичному рівні, проходить етап частотного аналізу, в результаті чого текстовим даним відповідає матриця $M(D)$, рядки якої відображають речення, стовпці – стеми, а значення формуються як число входжень стеми в речення. Над отриманою таким чином матрицею виконується операція сингулярного розкладання порядку $m \times n$, що розкрита в рамках латентно-семантичного аналізу:

$$M(D) = \begin{vmatrix} w_{1,1}(t_1) & & w_{1,n}(t_n) \\ \vdots & \cdots & \vdots \\ w_{m,1}(t_1) & & w_{m,n}(t_n) \end{vmatrix}, M(D) = U \sum V, \quad (5)$$

де n – кількість термінів, m – кількість речень, $w_{m,n}(t_n)$ – число входжень терміну n в речення m; $\sum$ – матриця розміру $m \times n$ з невід'ємними елементами, у якій елементи, що лежать на головній діагоналі – це сингулярні числа (а всі елементи, що не лежать на головній діагоналі, є нульовими), U (порядку m) і V (порядку n) – це дві унітарні матриці, що складаються з лівих і правих сингулярних векторів відповідно.

Оскільки сингулярне розкладання є стійким, стає можливим прибрати значення лівої і правої матриці, які відповідають низьким сингулярним значенням, залишивши тільки два найбільших, що представляють собою координати, які проектуються на двомірну площину як масив точок для стем і для речень.

Після завершення латентно-семантичного аналізу і проектування даних на площину, проводиться кластеризація отриманих точок-координат для стем і для речень за алгоритмом k-means. Головними особливостями алгоритму k-means є те, що, по-перше, необхідно заздалегідь знати кількість кластерів, а по-друге алгоритм дуже чутливий до вибору початкових центрів кластерів – класичний варіант має на

увазі випадковий вибір кластерів, що дуже часто є джерелом похибки. Тому в рамках розробки етапу семантичного аналізу були створені методи для встановлення необхідних параметрів кластеризації. Кількість кластерів для стем і для речень *cl* визначається за формулою:

$$cl(W, W_U) = \frac{count(W)}{count(W_U)}, \tag{6}$$

де $count(W)$ – кількість слів, $count(W_U)$ – кількість стем.

Центроїдами кластерів-стем є координати стем з найбільшою кількістю входжень в текст, які визначаються за формулою:

$$C_{st}(W_U) = \max(W_0 \cdots W_{cl}), \tag{7}$$

де $W_0 \ldots W_{cl}$ – кількості входжень стем із кластеру у текст.

Центроїдами кластерів-речень є координати речень з найбільшою загальною вагою стем, які визначаються за формулою:

$$C_S(W_S) = \max\left(\sum_{i=0}^{SN} W_i\right), \tag{8}$$

де W_i – кількість входжень стеми із речення у текст, *SN* – кількість стем в реченні.

В якості міри відстаней між об’єктами кластера була вибрана манхеттенська відстань.

Останньою операцією семантичного рівня є формування семантичних контурів у двомірній площині. Для цього, на основі координат точок кожного кластера-стеми будується контур опуклої фігури за алгоритмом Джарвіса, суть якого полягає у наступному: нехай на площині задана кінцева множина точок *A*. Оболонкою цієї множини називається будь-яка замкнута лінія *H* без самоперетинів така, що всі точки з *A* лежать всередині цієї кривої. Виходячи із вищесказаного, семантичним контуром кластеру називається опукла фігура, що була сформована для множини точок із одного кластеру-стеми або кластеру-речення. Отримані результати для кластерів-стем і для кластерів-речень представлені на рис. 2, де кожному кольору точки відповідає кластер, отриманий за алгоритмом k-means, а лінії обмежують приблизний контур опуклої фігури, отриманої за алгоритмом Джарвіса.

Заключною операцією побудови семантичної моделі тексту є рівень виведення, першим кроком якого є визначення кількості координат-стем C^S, що входять в кожний кластер-стему, після чого будується семантичний граф зв’язків кластерів-стем в порядку убування C^S, який є основою для побудови семантичної моделі. Для кожної фігури кластерів-стем, отриманої за алгоритмом Джарвіса, перевіряється попадання точок, які формують кожен кластер-речення. Якщо такі

точки знайдені – кластер речення приєднується в моделі до кластеру-стеми, де вага зв'язку – це кількість точок, що потрапили в контур кластера-стеми.

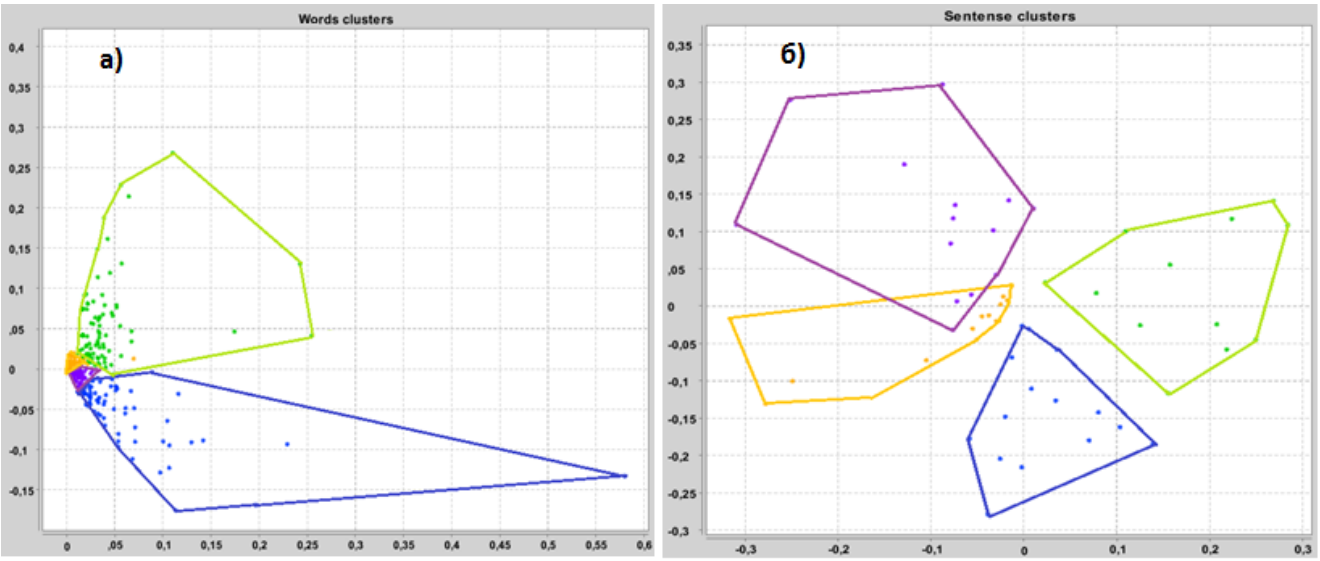

Рис. 2. Опуклі контори а) кластерів-стем, б) кластерів-речень

Отримана таким чином семантична структура M зберігається в базу знань і представлена сукупністю множин термінів T і сукупністю множин речень S, зв'язок між якими визначається матрицею семантичних ваг W:

$$T = \left\{ T_1\{t_1 \dots t_{w_{T_1}}\} \cdots T_n\{t_n \dots t_{w_{T_n}}\} \right\},$$
$$S = \{S_1\{s_1 \dots s_{l_1}\}\} \cdots S_1\{s_1 \dots s_{l_n}\}\}, \qquad (9)$$
$$W = \begin{vmatrix} \{w_{T1}, w_{S1}\}, & \{w_{T2}, w_{S1}\} & \cdots & \{w_{Tn}, w_{S1}\} \\ \vdots & \vdots & & \vdots \\ \{w_{T1}, w_{Sm}\}, & \{w_{T2}, w_{Sm}\} & \cdots & \{w_{Tn}, w_{Sm}\} \end{vmatrix},$$

де $t_1 \dots t_{w_{T,}}$ – множина координат стем, що формують кластер-стему, $s_1 \dots s_{S_l}$ – множина координат речень, що формують кластер-речення, $w_{T,}$ – кількість точок у кластері-стемі, S_l – розмір кластера речення, w_S – вага зв'язку між кластером-стемою і кластером-реченням, n – кількість кластерів-стем, $m = n$ – кількість кластерів-речень.

Графічний результат роботи етапу виведення зображений на рис. 3. Результуюча семантична модель складеться із кластерів-стем (позначено на рисунку як WordCluster), до яких прив'язане значення C^S (зображене на рисунку у квадратних дужках), та кластерів-речень (позначено на рисунку як SentenceCluster) які пов'язані із кластерами-стемами семантичним відношенням, вага якого позначена у круглих дужках.

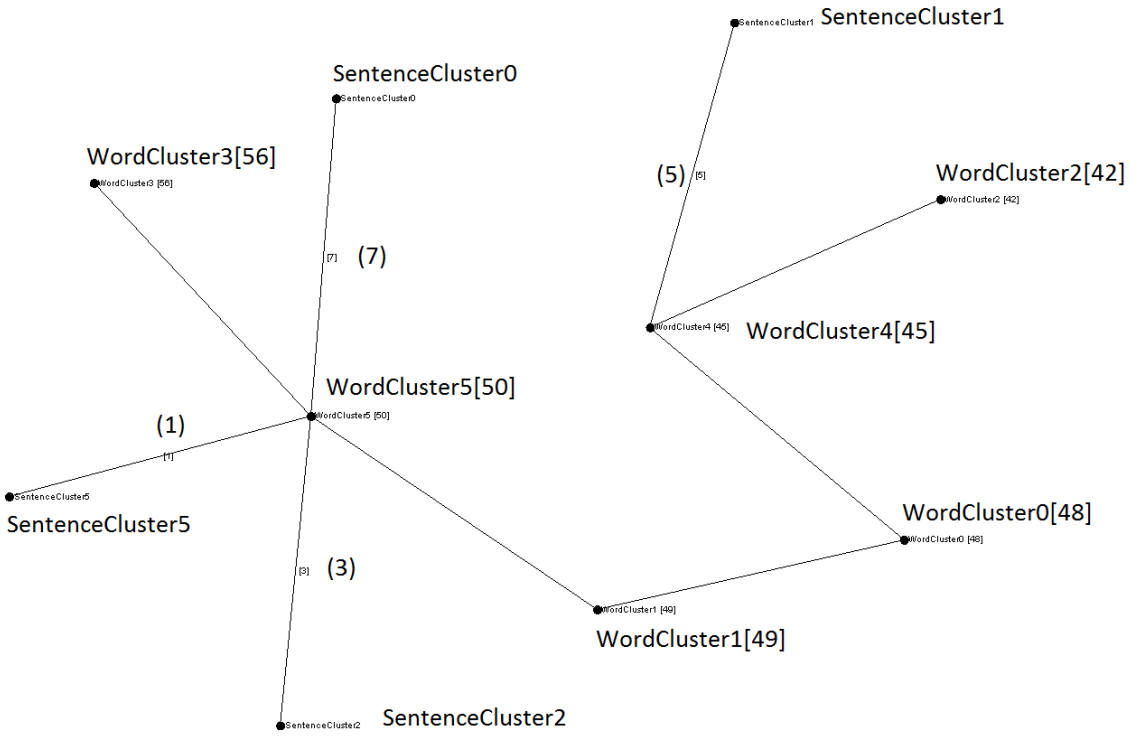

Рис. 3. Семантична модель документа

Здійснено тестування отриманої семантичної моделі і розробленої на її основі комп'ютерної системи. Найбільш показовою перевіркою моделі є обробка тексту, побудованого за допомогою автоматичного генератора. Для оцінки роботи моделей на згнерованих текстах було виконано попарне порівняння числових значень семантичних моделей автоматично згенерованого тексту і знання щодо значення середнього коефіцієнта семантичної значущості моделі μ:

$$\mu(D) = \frac{\sum_{i=0}^{N} \frac{f_{\min}(D_i, D_{i+1})}{f_{\max}(D_i, D_{i+1})}}{N},$$

$$f_{\min}(D_1, D_2) = \min\left\{\frac{\sum w_{SD_1}}{N_{D_1}}, \frac{\sum w_{SD_2}}{N_{D_2}}\right\}, \quad (10)$$

$$f_{\max}(D_1, D_2) = \max\left\{\frac{\sum w_{SD_1}}{N_{D_1}}, \frac{\sum w_{SD_2}}{N_{D_2}}\right\},$$

де $\sum w_{SD}$ – вага всіх семантичних зв'язків в моделі знання D, N_D – кількість кластерів в моделі знання D, $f_{\min}(D_1, D_2), f_{\max}(D_1, D_2)$ – функції, які повертають мінімальне і максимальне значення відносин, N – кількість експериментів.

Дане значення характеризує розподіл семантичних міток документа щодо семантичних зв'язків, і показує залежність цього значення для пар текстів в корпусі знань $D = \{d_1 \dots d_N\}$. Коефіцієнт семантичної значущості моделі був розрахований

для $D_g -$ корпусу автоматично створених документів, D_I – корпусу знань і D_{gI} – корпусу, в якому зіставляється згенерований текст і знання між собою. Дослідження показали, що зміни коефіцієнта семантичної значущості для корпусу згенерованого тексту і корпусу знань при порівнянні текстів однакових розмірів практично не відбувається: $\mu_g(D_g) = 0{,}911$ та $\mu_I(D_I) = 0{,}932$ відповідно. Однак, якщо порівнювати між собою знання і згенерований текст, очевидне значне падіння значення коефіцієнта до $\mu_{gI}(D_{gI}) = 0{,}335$. Це відбувається через те, що на відміну від згенерованого тексту, семантичні мітки знання мають значно більшу вагу і відповідно об'єднують більшу кількість семантично пов'язаних термінів, підвищуючи таким чином семантичну значущість моделі. Через значно більшу кількість кластерів-стем в моделях згенерованого тексту, підтверджується головна ідея використання моделі в задачах автоматичної обробки знань в системах генерації відповідей – терміни не об'єднуються в семантичні мітки, оскільки не існує ніякого семантичного зв'язку знань в початковому тексті, що наочно показує адекватність розробленої моделі.

Окремою і важливою перевіркою адекватності алгоритму побудови семантичної моделі тексту є оцінка сенсової місткості семантичних міток документу. Для оцінки значущості семантичних міток документу була висунута гіпотеза, яка має на увазі, що утворені семантичні мітки документа мають найбільшу кількість перетинів із семантичними контурами тоді, коли вони містять найбільшу кількість семантично значних стем у своєму складі. Для перевірки гіпотези було побудовано 65 семантичних моделей, для яких було розраховано коефіцієнт середньої семантичної ємності ε по наступним темам – економіка, філософія, інформаційні технології і астрономія:

$$\varepsilon = \frac{\sum_{i=0}^{N} \frac{N_H}{N_L}}{N}, \tag{11}$$

де N_H – кількість термінів, семантично пов'язаних з тематикою знання в кластері-стемі з найбільшою загальною вагою зв'язків з кластерами-реченнями (сильні кластери), N_H – кількість термінів, семантично пов'язаних з тематикою знання в кластері-стемі з нульовою загальною вагою зв'язків з кластерами-реченнями (слабкі кластери), N – загальна кількість проведених тестів для заданої теми.

Даний коефіцієнт показує, у скільки разів щільність семантично важливих термінів в сильних кластерах перевищує щільність семантично важливих термінів в слабких кластерах. Отримані результати розрахунку сенсової місткості для кожного тематичного набору текстів показують, що у проведених тестах кількість семантично значних термінів у семантично сильних кластерах значно перевищує кількість термінів у семантично слабких кластерах незалежно від кількості, розміру та тематичної спрямованості документів – для тематики «економіка» (15 документів) кількість термінів у сильних кластерах у середньому у 2,98 рази більша ніж у слабких, для тематики філософія (21 документ) – у 3,37 рази, для тематики «астрономія» (24

документа) – у 3,47 рази, для тематики «інформаційні технології» (5 документів) – у 4,44 рази.

У третьому розділі розроблено модель пошукової системи, суть якої можна сформулювати наступним чином. Нехай $D = \{d_1, d_2 \ldots d_n\}$ – множина знань, розміром n, що складається з текстів $\{d_1, d_2 \ldots d_n\}$, T – множина термінів, що являє собою запит користувача. Тоді, множина відповідей системи представлена матрицею M (14):

$$M = \begin{vmatrix} f_{1,1}(t_1, d_1) & & f_{1,n}(t_n, d_1) \\ \vdots & \cdots & \vdots \\ f_{m,1}(t_1, d_m) & & f_{m,n}(t_n, d_m) \end{vmatrix},$$

$$f(t,d) = \begin{cases} 1, t \in T \wedge t \in V(d) \\ 0, t \notin T \wedge t \notin V(d) \end{cases}, \quad (12)$$

$$T \subset \{t_1 \ldots t_n\},$$

де $V(d)$ – множина сильних кластерів-стем для документа d – таких, що мають найбільшу загальну вагу зв'язків із кластерами речень, $f(t,d)$ – повертає 1, якщо термін t належить множинам T та $V(d)$, або 0 у іншому випадку.

Наведено опис головних сучасних теоретичних парадигм реалізації пошукових моделей «запит-відповідь», розкриті структура і алгоритм роботи створеної моделі пошукової системи із використанням генерації текстів, показані архітектури програмного додатку і створеної неструктурованої бази знань, а також проведене тестування моделі генерації відповідей на основі неструктурованої бази знань в пошукових системах і перевірена її адекватність. Наведено приклади отриманих за допомогою системи відповідей та проведені їх аналіз і оцінка.

Розкрито один з головних компонентів моделі генерації текстів – неструктурована база знань, наповнення якої прямим чином впливає на якість одержуваних результатів. Сформульоване поняття корпусу текстів та його основні характеристики. Розкриті структура та архітектура створеної бази даних, визначені атрибути кожної з таблиць. Встановлена відповідність створеної бази даних і головних характеристик корпусу. Показане відсоткове розподілення типів тексту у колекції для підтримки властивості репрезентативності корпусу. Зазначено, що оскільки в структурі розробленої бази даних міститься, окрім неструктурованих текстів, різноманітна інформація про структуру семантичної моделі, що являє собою правила виводу і генерації нових знань при здійснені пошуку відповіді на запит користувача, стає можливим стверджувати, що створена модель працює саме з базою знань неструктурованих текстів.

Розкрите питання застосування процесу попередньої оцінки і автоматичної фільтрації текстових документів за їх семантичною зв'язністю для запобігання наповнення системи даними, що є недостатньо якісними для її стабільної роботи. Сформульована задача класифікації текстів за їх семантичними властивостями та

оцінені актуальність і доцільність використання семантичної моделі для вирішення задачі фільтрації текстів. За створеною моделлю, кожен з текстів характеризується двома значеннями – нормалізованим розміром тексту W_N, отриманим за формулою:

$$W_N = \frac{W_I - W_{\min}}{W_{\max} - W_{\min}}, \tag{13}$$

де W_i – загальна кількість слів, W_{min}та W_{max} – найменша та найбільша кількість слів у навчальному корпусі, та нормалізованим семантичним значенням S_N, отриманим за формулою:

$$S_N = \frac{W_U}{W} * \frac{CW_C}{CW}, \tag{14}$$

де W – загальна кількість слів, W_U – загальна кількість стем, CW_C – кількість кластерів – стем, що мають зв'язок із кластерами-реченнями, CW – загальна кількість кластерів – стем.

Отримані таким чином данні є критеріями оцінки семантичної зв'язності тексту і водночас навчальною вибіркою для нейронної мережі. Розкрито структуру використаної нейронної мережі, процес її навчання для вирішення цільової задачі та проведені тести адекватності моделі, які показали достатню точність автоматичної класифікації, що вказує на коректну фільтрацію текстових даних.

Розглянуті основні компоненти програмної архітектури створеної системи, деякі зовнішні бібліотеки, що застосовані у процесі розробки та особливості функціоналу інтерфейсу додатку. Розроблено та впроваджено систему юніт- та інтеграційних тестів функціоналу системи, яка має забезпечити процес безпечної модифікації існуючого функціоналу. Проведений аналіз розробленої програмної архітектури показав, що основні концепції, яким необхідно слідувати при реалізації текстового автомату, виконані: властивість модульності реалізована за допомогою розбиття інструментів аналізу та обробки текстових знань на окремі програмні об'єднання; властивість рівневості реалізована завдяки розділенню функцій системи на окремі методи та завдяки лінійному алгоритму роботи системи, що реалізує покрокове застосування цих методів; властивість орієнтованості на використання бази знань забезпечується використанням колекції текстів, супутньої бази даних та програмними інструментами роботи з нею; безпека роботи системи забезпечується виконанням попередньої фільтрації знань та набором автоматизованих тестів результатів роботи системи.

Розроблено алгоритм тестування результатів роботи моделі, за яким проведена перевірка адекватності роботи пошукової системи «запит-відповідь». Сенс створеної перевірки виходить з ідеї про те, що відповіді системи, за умови роботи із адекватною базою знань, не можуть бути однозначно визначені як «правильні» або «неправильні», оскільки навіть фрагмент семантично зв'язного текстового знання містить у собі деяку кількість корисної для користувача інформації. Тому, у нашому

випадку, критерієм семантичної зв’язності відповіді служить тематична відповідність фрагменту знання, що є кандидатом до включення у результуючу відповідь, та набору термінів із запиту користувача. Виходячи з цього, розроблений алгоритм роботи системи тестування побудований на оцінці залежностей між тематикою множин знань, що є кандидатами до включення у результуючу відповідь, та тематикою множин автоматично сформованих до моделі запитів, що заздалегідь були назначені кожному документу у колекції текстів. Розглянуто основні кроки алгоритму роботи системи тестування та два режими функціонування пошукової системи: семантично відповідний та семантично невідповідний. Зазначено, що для розуміння ступеню впливу семантичної моделі на процес генерації тексту необхідно розрахувати коефіцієнт коректності формування множини документів Q_D за формулою:

$$Q_D = \frac{\sum N_C}{\sum N_W}, \tag{15}$$

де $\sum N_C$ – загальна кількість ситуативно коректних документів для кожного запиту, $\sum N_W$ – загальна кількість ситуативно некоректних документів для кожного запиту.

Гіпотеза перевірки полягає у тому, що значення коефіцієнту коректності формування множини документів для семантично відповідного режиму має бути більшим, ніж для семантично невідповідного. Для її підтвердження було автоматично сформовано і виконано близько 1000 запитів до системи, у результаті чого було сформовано множини текстових знань на основі 20 тисяч документів, а значення коефіцієнту коректності формування множини документів для семантично відповідного режиму склало 0,843 проти значення у 0,493 для семантично невідповідного режиму, що вказує на доцільність застосування підходу заснованого на семантичних моделях у системі генерації тексту відповіді на поставлене питання.

Заключним кроком перевірки адекватності моделі є мануальне тестування, яке базується на індивідуальних оцінках відповідей системи за бальним методом і являє собою сукупність оцінок вимог до згенерованої відповіді від 0 до 1 із кроком 0,1. Усього таких вимог 5, тобто кожна відповідь сумарно може получити від 0 до 5 балів. У ході дослідження був проведений ряд тестів із різною складністю питань і різним тематичним напрямом запитів, на кожен з яких була отримана відповідна оцінка. Середнє значення усіх проведених експертних оцінок склало 0,839, що вказує на задовільні результати роботи системи.

ВИСНОВКИ

В дисертаційній роботі розв’язано важливу науково-технічну задачу обробки семантично-неструктурованих документів для генерації відповідей в пошукових системах. Отримані нові обґрунтовані результати, які відповідно до поставленої мети дають вирішення актуальної задачі побудови моделі отримання знань із неструктурованих джерел.

Основні наукові та практичні результати полягають в наступному:

1) Проведено аналіз існуючих підходів до побудови математичної моделі представлення мови та вилучення інформації з неструктурованої бази знань. За його результатами зроблено висновок про доцільність використання моделі м'якого розуміння Леонтьевої у основі процесів розуміння семантики тексту і генерації відповідей у пошукових системах. До переваг обраної моделі відносяться орієнтованість на складні системи інтелектуальної оброки текстових даних та на першочергову обробку семантичних характеристик знань.
2) Отримала подальший розвиток теорія «Сенс-Текст» та модель м'якого розуміння знань у пошуковій системі вилучення інформації із неструктурованої бази знань через подолання обмежень тлумачно-комбінаторного словника, використання якого вимагає глибокого ручного опису кожного компонента текстової бази знань, що є складною глобальною задачею прикладного застосування цих моделей.
3) Розроблено математичну модель текстового автомату пошукової системи, дотримання концепцій якого дозволило описати програмну архітектуру системи як комплекс послідовних рівнів обробки бази знань. Розроблені та реалізовані головні критерії до функціонування системи, що забезпечило її архітектурну консистентність та обґрунтованість.
4) Розроблено семантичну модель текстового знання, яка використовує у своїй роботі як базові методи синтаксичної обробки тексту, так і науково нові застосування латентно-семантичного аналізу, що у комплексі дозволило створити семантичну модель без залучення додаткових лінгвістичних знань.
5) Реалізовано семантичну модель текстового знання, яка представлена у вигляді програмного API із відкритим доступом, що дозволяє розробникам комп'ютерних систем отримувати кількісні значення семантичних характеристик текстових даних без необхідності залучення лінгвістичних знань.
6) Розроблено математичну модель автоматичної класифікації документів за їх семантичною зв'язністю, яка дозволила забезпечити надійність процесу наповнення неструктурованої бази знань новими знаннями.
7) Створено математичну модель генерації відповідей на основі неструктурованої бази знань в пошукових системах. Розроблено комп'ютерну систему, яка реалізує створену модель генерації відповідей і дозволяє отримувати на основі семантичних моделей неструктурованих знань нові знання, що містять відповіді на поставлені користувачем питання.
8) Розроблено систему оцінювання якості моделі генерації відповідей на основі неструктурованої бази знань та пошукової системи. Проведене дослідження показує доцільність використання процесу генерації текстів для формування відповіді на поставлене питання. Отримані результати проведеного автоматичного і мануального тестування системи вказують на коректність і адекватність роботи додатку.

Розроблена комп'ютерна система впроваджена у міський комунальний заклад культури «Централізована система бібліотек для дітей» м. Дніпро як пошуковий інструмент обробки електронних текстів, у ТОВ «Сітал Україна» як засіб автоматичної генерації текстових інструкцій та у АТ «ДніпроАзот» як інструмент покращення процесів пошуку в системах електронного документообігу. Результати дисертаційної роботи опубліковані в 14 наукових працях, в тому числі 8 статей – у журналах, рекомендованих МОН України для публікації результатів дисертацій, та закордонних виданнях, та 6 – у тезах доповідей та трудах міжнародних та всеукраїнських конференцій.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні результати дисертації:

1. Волковський О.С. Computer methods for compiling an entry of explanatory combinatorial dictionary belonging to «Meaning↔Text» theory within the task of text automatic generation / О.С. Волковський, Є.Р. Ковилін // Математичне моделювання // науковий журнал – Кам'янське, 2018. – №2 (39). – с. 9–18. *Видання включено до НМБ Index Copernicus International.*
2. Волковський О.С. Analysis of the modern approaches to the problem of the automatic text generation in the natural language / О.С. Волковський, Є.Р. Ковилін // Системні технології. Регіональний збірник міжвузівських наукових праць // науковий журнал – Дніпро, 2016. – №5 (106). – с. 3–12. *Видання включено до НМБ Index Copernicus International*.
3. Волковський О.С. Комп'ютерна система автоматичного визначення зв'язності тексту / О.С. Волковський, Є.Р. Ковилін // Системні технології. Регіональний збірник міжвузівських наукових праць. // науковий журнал – Дніпро, 2017. – №1 (112). – с. 11–17. *Видання включено до НМБ Index Copernicus International.*.
4. Волковський О.С. Комп'ютерна система автоматичного аналізу промислових інструкцій. / О.С. Волковський, Є.Р. Ковилін // Системні технології. Регіональний збірник міжвузівських наукових праць. // науковий журнал – Дніпро, 2018. – №3 (116). – с. 28–37. *Видання включено до НМБ Index Copernicus International*.
5. Волковський О.С. Комп'ютерна система інтелектуального семантичного пошуку з використанням генерації текстів / О.С. Волковський, Є.Р. Ковилін // Вісник Херсонського національного технічного університету // науковий журнал – Херсон, 2018. – №3 (66). – с. 238 –245. *Видання включено до НМБ Google Scholar*.
6. Волковський О.С. Mathematical model for automatic creation the semantic thesaurus for the scientific text / О.С. Волковський, Є.Р. Ковилін // Системні технології. Регіональний збірник міжвузівських наукових праць. // науковий журнал – Дніпро, 2019. –. №6 (125). – с. 82–88. *Видання включено до НМБ Index Copernicus International*.

7. Волковський О.С. Модель автоматичної оцінки адекватності комп'ютерних систем «запит-відповідь» з використанням генерації текстів / О.С. Волковський, Є.Р. Ковилін // Системні технології. Регіональний збірник міжвузівських наукових праць. // науковий журнал – Дніпро, 2020 – №4 (129). – с. 50–58. *Видання включено до НМБ Index Copernicus International*.
8. Volkovsky O.S. Matematical model for constructing the semantic network of a scientific text / O.S. Volkovsky, Y. R. Kovylin. // Modern engineering and innovative technologies // науковий журнал – Карлсруе, Німеччина, 2020 – №11 – с. 128 – 133. *Видання включено до НМБ Index Copernicus International*.

Публікації апробаційного характеру:

9. O.S. Volkovsky. Computer System of Building of the Semantic Model of the Document / O.S. Volkovsky, Y. R. Kovylin. // 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP): міжнародна науково-практична конференція, 21-25 серпня 2018 р.: тези доп. – Львів, 2018 – с. 322–327. **(*Конференція включена до НМБ Scopus*).**
10. Волковський О.С. Система автоматичного аналізу текстів природною мовою / О.С. Волковський, Є.Р. Ковилін // Інформаційні технології в металургії та машинобудуванні: міжнародна науково-технічна конференція, 28–30 березня 2017р.: тези доп. – Дніпро, 2017 – с. 112.
11. Волковський О.С. Автоматична побудова семантичної мережі тексту у системах запит-відповідь / О.С. Волковський, Є.Р. Ковилін // Інформаційні технології та комп'ютерне моделювання: міжнародна науково-практична конференція, 15 – 20 травня 2017 р.: тези доп. – Івано-Франківськ, 2017– с. 386 – 389.
12. Волковський О.С. Комп'ютерна система автоматичного аналізу промислових інструкцій.. / О.С. Волковський, Є.Р. Ковилін // Інформаційні технології в металургії та машинобудуванні: міжнародна науково-технічна конференція, 27–29 березня 2018 р.: тези доп. – Дніпро, 2018. – с. 124.
13. Волковський О.С. Комп'ютерна модель семантичної мережі документу в системі запит-відповідь / О.С. Волковський, Є.Р. Ковилін // Проблеми математичного моделювання: всеукраїнська науково-методична конференція, 23–25 травня 2018 р.: тези доп. – Кам'янське, 2018. – с.33–36.
14. Волковський О.С. Комп'ютерна система інтелектуального семантичного пошуку з використанням генерації текстів / О.С. Волковський, Є.Р. Ковилін // Матеріали XIX міжнародної конференції з математичного моделювання, присвяченої 250-річчю з дня народження Жозефа Фур'є, 17–21 вересня 2018 р.: тези доп. – Лазурне, 2018. – с. 49.

АНОТАЦІЯ

Ковилін Е.Р. Модель генерації відповідей в пошукових системах на основі неструктурованої бази знань – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 01.05.02 – математичне моделювання та обчислювальні методи. – Національна металургійна академія України, Дніпро 2020.

Дисертаційну роботу присвячено вирішенню актуальної науково-прикладної задачі розробки моделі для автоматизації обробки семантично-неструктурованих документів для генерації відповідей в пошукових системах.

На основі розроблених математичних моделей автоматичного отримання семантичних характеристик тексту і автоматичної класифікації вхідних даних за ступенем їх семантичної зв'язності, запропоновано модель і алгоритм генерації відповідей на запит користувача на основі неструктурованої бази знань.

У вигляді відкритого програмного API реалізовані: семантичнаа модель тексту, модель автоматичної класифікації тестів за ступенем їх зв'язності та модель автоматичної генерації текстів на основі неструктурованої бази знань.

Використання отриманих у роботі результатів дозволяє отримувати семантичні моделі текстів без залучення ручної семантичної розмітки або лінгвістичних знань і спростити роботу користувача із пошуковими системами.

Ключові слова: математична модель пошукової системи, семантична модель, неструктуровані текстові знання, генерація відповідей, комп'ютерні моделі представлення знань.

ABSTRACT

Kovylin Y. R. Model for generating answers based on an unstructured knowledge base in the search systems. - Manuscript.

Thesis for obtaining the candidate degree (Ph.D) in engineering sciences in the specialty 01.05.02 – Mathematical modeling and computational methods (Technical science). – Oles Honchar Dnipro National University, The National Metallurgical Academy of Ukraine, Dnipro, 2020.

The dissertation is devoted to the solution of the actual scientific and applied problem of model development for automation of processing the semantically unstructured documents for the answers generation in the search engines.

The dissertation analyzes the existing approaches to the construction of applied models of the texts in natural language. Existing approaches require the involvement of significant volumes of linguistic knowledge, preliminary construction of ontological markup or manual creation of the semantic dictionaries of knowledge, which significantly reduces the adaptability of models and complicates their applied use. The existing software systems rely in their work on the extraction of knowledge using ontological bases, which makes it impossible to use them for processing the unstructured knowledge bases.

Based on the results of the analysis, it became obvious that there is a need to develop the models that allows the automating process of obtaining and presenting semantic models of specialized texts without the need to involve linguistic knowledge or the formation of the preliminary semantic markups or dictionaries.

The first developed model was a mathematical model for obtaining the semantic characteristics of a specialized text. The created model is innovative, and uses in its work the method of latent semantic analysis, clustering and spatial analysis. The created model makes it possible to obtain a semantic model of a text without using a previously created semantic structure or semantic dictionaries, which made it possible to fully automate the process of obtaining quantitative values of the semantic characteristics of the text. The carried out tests showed that, despite the use of the frequency characteristics of the text, the created model depends precisely on the semantics of a document, and not on its frequency portrait. In addition, the created model makes it possible to reliably combine semantically related terms into semantic text labels, while establishing a semantic connection with a set of sentences from the source text, which is a prerequisite for building an automatic text generation process.

The second developed model was the model of automatic classification of incoming documents, which uses the quantitative characteristics from the semantic model of text in the process of its work. The created model allows system to filter the incoming documents, thus protecting the created knowledge base from being filled with inappropriate texts. The carried out tests showed that the accuracy of the system, built on the basis of the created model, is 90%, which is sufficient for the correct operation of this stage.

The last model created is the unstructured knowledge base response generation model. This model works on the basis of the first and second models and allows to generate the new text knowledge relevant to the user's request. The tests have shown that the use of the developed model allows improvements in the semantic quality of the set of candidate texts for generation by 1.7 times, in comparison with the direct solution of the search problem. The expert assessments carried out by the ball method showed a value of 0.839, which proves the applicability and adequacy of the created model.

The developed software application was implemented in the city municipal cultural institution «Centralized system of libraries for children» in Dnipro as a search tool for electronic text processing, in LLC «Sital Ukraine» as a tool of automatic generation of text instructions and in JSC «DniproAzot» as a tool to improve search processes. The results of the dissertation are published in 14 scientific papers, including 8 articles - in journals recommended by the Ministry of Education and Science of Ukraine for publication of dissertations and foreign publications, and 6 - in abstracts and papers of international and national conferences.

Keywords: mathematical model of search engine, semantic model, unstructured text knowledge, generation of answers, computer models of knowledge representation.