УДК 004.912

Є.Р. КОВИЛІН, О.С. ВОЛКОВСЬКИЙ
Дніпровський національний університет ім. Олеся Гончара

КОМП'ЮТЕРНА МОДЕЛЬ ГЕНЕРАЦІЇ ВІДПОВІДЕЙ У ПОШУКОВІЙ СИСТЕМІ НА ОСНОВІ НЕСТРУКТУРОВАНОЇ БАЗИ ЗНАНЬ

Метою роботи є розробка моделі системи запит-відповідь, що спроможна створювати конкретні текстові відповіді на запит користувача, використовуючи у своєму алгоритмі генерацію наукового тексту на природній мові. Система визначає смислові зв'язки в документах, створюючи при цьому новий текст, який містить відповідь на питання користувача. У статті розглядається модель системи, що базується на розробленому підході до формування семантичної моделі документа, який дозволяє отримувати кількісні показники семантичних властивостей документу на природній мові і сенсові зв'язки між компонентами тексту. Розроблена у вигляді прикладного програмного автомата, система семантичного пошуку має використовувати модель, спроможну працювати із достатньо формалізованим типом знань, а саме – науковим текстом і дозволяти автоматично формувати програмну семантичну модель як окремого документа, так і всього корпусу знань в цілому. На основі отриманої структури додаток має генерувати текстову відповідь на отриманий користувальницький запит. Це призводить до важливої наукової властивості створеної моделі – додаток повинен мати можливість використовувати нерозмічений заздалегідь корпус текстів, що являє собою неструктуровану базу знань, задля чого необхідно створити та дослідити семантичну модель наукового тексту на природній мові, а також розробити алгоритм її формування з семантичної мережі. Такий підхід вирішує більшість питань обробки тексту задля подальшої автоматичної генерації відповіді. Додатково розроблена підсистема автоматичної класифікації наукових текстів за ступенем їх зв'язності, що використовує у своїй роботі кількісні характеристики семантичних властивостей створеної моделі наукового тексту. У статті описані розроблені критерії оцінки створених систем та алгоритмів. Отримана таким чином система, окрім організації зручного пошукового середовища, утворює універсальну модель для проведення автоматичної обробки текстів на семантичному рівні для груп слов'яномовних текстів формального стилю, набір інструментів якої дозволяють гнучко створювати і оброблювати тематичні повнотекстові корпуси документів без попередньої семантичної розмітки та отримати програмну модель тексту формалізованої стильової спрямованості із кількісними характеристиками семантичних властивостей тексту, на основі яких можливо вирішувати інші завдання автоматичної обробки текстів.

Ключові слова: семантична мережа; автоматична обробка тексту; система запит-відповідь; генерація тексту.

Е.Р. КОВЫЛИН, О.С. ВОЛКОВСКИЙ
Днепровский национальный университет им. Олеся Гончара

КОМПЬЮТЕРНАЯ МОДЕЛЬ ГЕНЕРАЦИИ ОТВЕТОВ В ПОИСКОВОЙ СИСТЕМЕ НА ОСНОВЕ НЕСТРУКТУРИРОВАННОЙ БАЗЫ ЗНАНИЙ

Целью работы является разработка модели запросно-ответной системы, способной создавать конкретные текстовые ответы на запрос пользователя, используя в своем алгоритме генерацию научного текста на естественном языке.

10.32782/KNTU2618-0340/2020.3.2-1.13

Система определяет смысловые связи в документах, создавая при этом новый текст, содержащий ответ на вопрос пользователя. В статье рассматривается модель системы, основанной на разработанном подходе к формированию семантической модели документа, который позволяет получать количественные показатели семантических свойств документа на естественном языке и смысловые связи между компонентами научного текста. Разработанная в виде прикладного программного автомата, система семантического поиска должна использовать модель, способную работать с достаточно формализованным типом знаний, а именно - научным текстом и позволять автоматически формировать программную семантическую модель как отдельного документа, так и всего корпуса знаний в целом. На основе полученной структуры приложение генерирует текстовый ответ на полученный пользовательский запрос. Это приводит к важному научному свойству созданной модели - приложение должно использовать неразмеченный заранее корпус текстов, который представляет собой неструктурированную базу знаний, для чего необходимо создать и исследовать семантическую модель научного текста на естественном языке, а также разработать алгоритм ее формирования из семантической сети. Такой подход решает большинство вопросов обработки текста для дальнейшей автоматической генерации ответа. Дополнительно разработана подсистема автоматической классификации научных текстов по степени их связности, использующая в своей работе количественные характеристики семантических свойств созданной модели научного текста. В статье описаны разработанные критерии оценки созданных моделей и алгоритмов. Полученная таким образом система, кроме организации удобной поискового среды, образует универсальную модель для проведения автоматической обработки текстов на семантическом уровне для групп славяноязычных текстов формального стиля, набор инструментов которой позволяют гибко создавать и обрабатывать тематические полнотекстовые корпуса документов без предварительной семантической разметки и получить программную модель текста формализованной стилевой направленности с количественными характеристиками семантических свойств текста, на основе которых возможно решать другие задачи автоматической обработки текстов.

Ключевые слова: семантическая сеть; автоматическая обработка текста; система запрос-ответ; генерация текста.

Y.R. KOVYLIN, O.S. VOLKOVSKY
Oles Gonchar Dnipro National University

COMPUTER MODEL OF RESPONSE GENERATION IN THE SEARCH SYSTEM BASED ON AN UNSTRUCTED KNOWLEDGE BASE

The aim of the work is to develop a request-response system model capable of creating specific textual responses to a user's request, using a scientific text generating in a natural language in its algorithm. The system determines the semantic links in the documents, while creating a new text containing the answer to the user's question. The article discusses a model of a system based on the developed approach to the formation of a semantic model of a document, which allows you to get quantitative indicators of the semantic properties of a document in a natural language and semantic links between components of a scientific text. Developed as an application software, the semantic search system should use a model capable of working with a sufficiently formalized type of knowledge, namely - scientific text and allow you to automatically form a software semantic model of a single document and the body of knowledge as a whole. Based on the received structure, the application should

10.32782/KNTU2618-0340/2020.3.2-1.13

generate a text response to the received user request. This leads to an important scientific property of the created model - the application should be able to use unstructured corpus of texts, which is an unstructured knowledge base, for which it is necessary to create and explore a semantic model of scientific text in natural language, and develop an algorithm for its formation from the semantic network. This approach solves most word processing issues for further automatic generation. In addition, a subsystem for the automatic classification of scientific texts by the degree of their connectivity was developed, which uses quantitative characteristics of the created model of a scientific text in its work. The article describes the developed criteria for evaluating the created systems and algorithms. The system thus obtained, in addition to organizing a convenient search environment, forms a universal model for automatic text processing at a semantic level for groups of Slavic-language texts of a formal style, a set of tools that allow you to flexibly create and process thematic full-text document bodies without preliminary semantic markup and get a program text model formalized stylistic orientation with quantitative characteristics of semantic properties text and, on the basis of which it is possible to solve other problems of automatic word processing.

Keywords: semantic web; automatic text processing; request-response system; text generation.

Постановка проблеми

Проблема оптимального шляху пошуку інформації є однією з ключових в області комп'ютерної науки. Процес розробки більшості програмних продуктів рано чи пізно призводить до необхідності реалізації механізмів додавання, збереження і отримання інформації для її подальшої обробки. Ефективним і головним наразі рішенням є різноманітні системи баз даних, що прекрасно справляються з цими завданнями на програмному рівні. Однак, якщо в функціоналі системи присутня необхідність працювати безпосередньо з призначеним для користувача запитом, який часто складається з декількох неформальних критеріїв і вимагає певного семантичного аналізу, то обробка отриманих результатів повністю лягає на плечі користувача. Йдеться про підхід, який використовується у багатьох популярних web-пошукових системах: відповіддю на отриманий запит є множина ранжованих гіпертекстових документів (web-сторінок), 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, через що в отриманому масиві документів міститься велика кількість не пов'язаної із запитом користувача інформації, а також множини повторень однакової інформації, поданої в різних інтерпретаціях. Представлена робота присвячена розробки системи запит-відповідь, що спроможна створювати конкретні текстові відповіді на запит користувача, використовуючи у своєму алгоритмі генерацію тексту на природній мові. Особливістю процесу прикладної розробки таких систем є необхідність вирішення великої кількості наукових проблем галузі і залучення провідних інструментів штучного інтелекту для вирішення цільової задачі. Складності додає і сама структура мови – підходи, що використовуються для обробки однієї мови, можуть не спрацювати для іншої, наприклад через явище флексії (як для української та англійської мов). Головною ж проблемою є необхідність попереднього ручного опису семантичних онтологій між усіма елементами мови і між усіма документами в системі, що є важкою глобальною задачею галузі.

Аналіз останніх досліджень і публікацій

Проведені дослідження існуючих розробок систем «запит-відповідь», що формують безпосередню відповідь на запит користувача, дозволяють виділити два

10.32782/KNTU2618-0340/2020.3.2-1.13

основних напрями розробки таких систем - на основі лексико-семантичного словника відносин і на основі статистичного аналізу текстів. Представником першого напряму є система [1], де семантичний аналіз полягає у виявленні взаємозв'язків між об'єктами (персоналіями, організаціями, подіями) і класифікації відносин між ними, а також ототожненні об'єктів із заздалегідь заданими семантичними класами, а другого напряму - система [2], де для вхідного запиту складається "інформаційний портрет" - набір упорядкованих за значимістю ключових слів і словосполучень, характерних саме для даної вибірки текстів, після чого за набором ключових слів користувач може самостійно визначити теми, які можуть бути видані у відповідь на його запит, і тим самим уточнити потрібну йому тематику. Описані підходи не вирішують порушених в роботі проблем, оскільки мають на увазі залучення великої кількості ручної праці, як з боку користувача, при аналізі отриманих результатів, так і з боку розробника, при складанні попередньої семантичної розмітки, що суперечить поставленій в цій роботі меті.

Розроблена система базується на створені семантичної моделі тексту, тому додатково було проведено аналіз основних підходів до комп'ютерного формування семантичних моделей тексту як для слов'янських, так і для англійської мов. Встановлено, що основою всіх цих підходів, що формують базові відносини між елементами в тексті, є продукційна модель онтологій. Практичне застосування цієї технології детально описано в роботі [3], на основі якої створюється семантичне уявлення метаописів тестового документа для подальшого семантичного пошуку. Важливою особливістю в рамках нашої роботи є те, що вихідні дані системи формуються на основі заздалегідь розміченого вручну корпусу мови. Цікавою практичною розробкою з використанням семантичних мереж є система формування семантичної мережі з слабоструктурованих текстових джерел, описана в роботі [4]. Автори роботи пропонують підхід для автоматичного відновлення структури розділів статті відкритого словника Wiktionary. Особливістю даного підходу є розробка деякої системи правил, на основі яких функціонує семантична програмна модель статті. Описані класи прикладних розробок комп'ютерних систем побудови семантичних мереж передбачають використання в якості вихідної бази знань деякий масив текстів, що містять попередню лінгвістичну розмітку, що, як було сказано вище, не задовольняє поставленій в роботі меті.

Мета дослідження

Рішення наукових проблем опису семантичних онтологій для корпусу наукових текстів на природній мові за допомогою розробки семантичної інтелектуальної системи запит-відповідь, що орієнтується під час своєї роботи на механізм генерації текстів. Система автоматично будує семантичну модель тексту, на основі якої визначає смислові зв'язки в документах, створюючи при цьому новий текст, який містить відповідь на питання користувача. Система не має використовувати у своїй роботі будь-які попередньо закладені семантичні знання про тексти, що оброблює.

Викладення основного матеріалу дослідження

Описана в цій статті модель є розвитком попередніх досліджень про побудову системи автоматичної генерації текстів на основі концепції моделі м’якого розуміння Леонтьєвої [5] та побудову семантичної моделі наукового тексту [6]. Розглянемо алгоритм роботи системи, що зображений на рис.1 та умовно поділений на п’ять основних кроків, більш детально.

Крок перший – отримання і обробка запиту користувача, що є типовою поведінкою систем запит-відповідь. У нашому випадку, під запитом розуміється набір

10.32782/KNTU2618-0340/2020.3.2-1.13

ключових слів, інформація про які цікавить користувача, розділених між собою знаками пробілу. Наприклад: «Класи програмування розробка»; «Космос комети»; «Телескопи»; «Теорія економіки». Форма та відмінок слів що складає запит не фіксовані, і може задаватися на бажання користувача. В цілому, система виходить з того, що семантична складова запиту є цільною і його частини не суперечать одна одному. Така форма запиту є відмінною від звичних морфологічно повних фраз, і була обрана для спрощення перевірки функціонування системи – аналіз повнотекстових запитів користувача є окремим семантико-морфологічним завданням і виходить за рамки цілей цієї роботи.

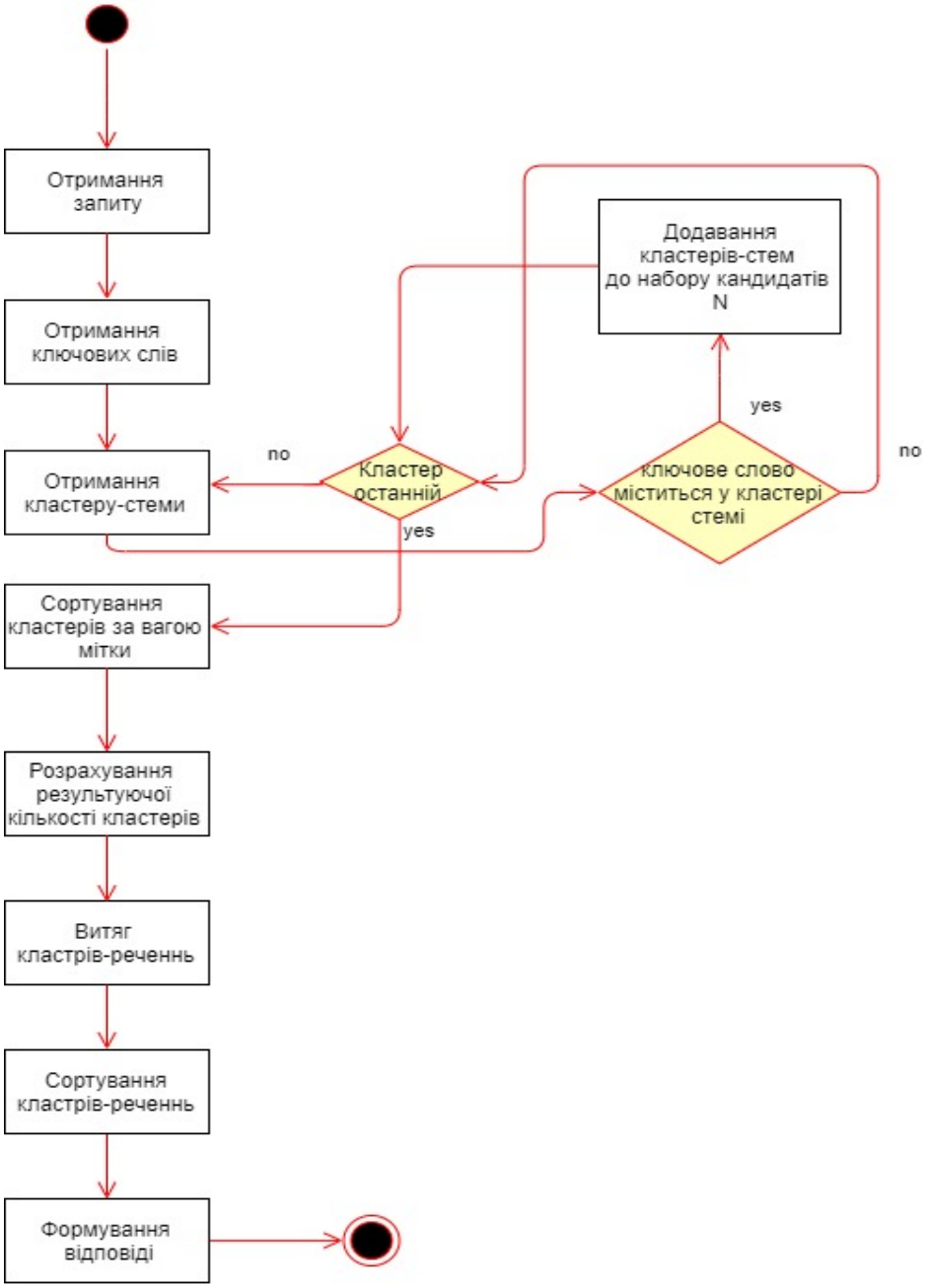

Рис. 1. Структура роботи системи запит-відповідь.

10.32782/KNTU2618-0340/2020.3.2-1.13

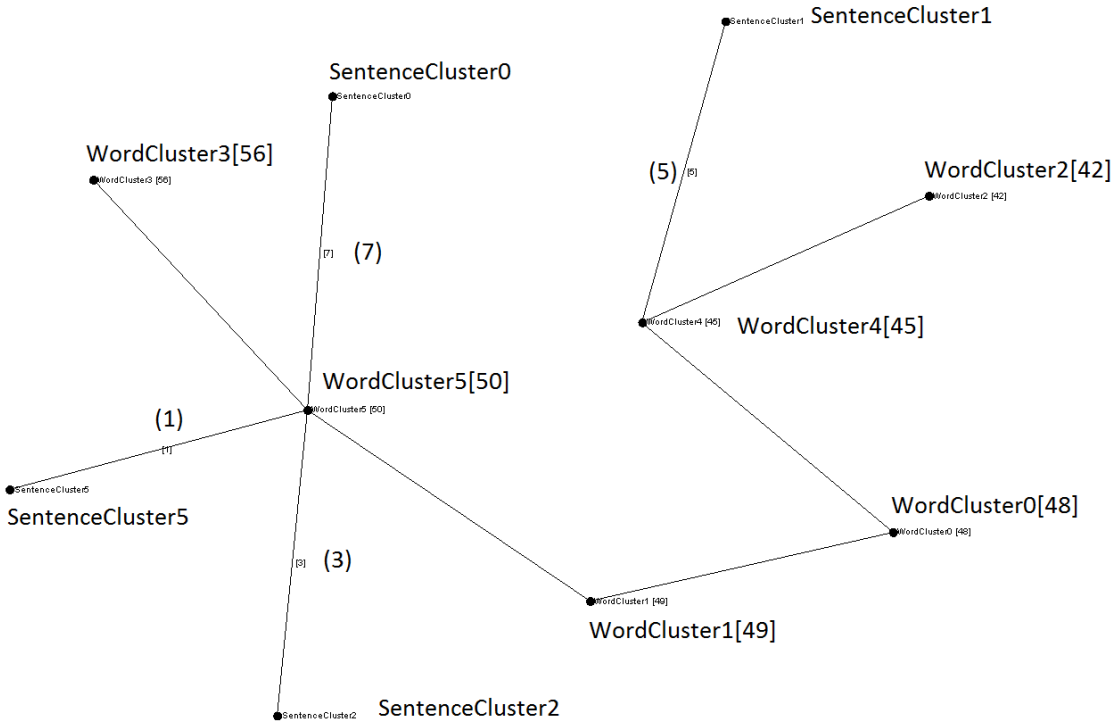

Рис. 2. Семантична мережа наукового тексту.

Крок другий – для кожного окремого ключового слова із запиту користувача відбувається операція пошуку відповідних кластерів-стем, множини яких формують семантичні мітки документа, за якими і буде генеруватися результуюча відповідь. Алгоритм побудови семантичної моделі документа і виділення кластерів-стем був описаний у роботі [6] і є інноваційною розробкою створеною в рамках цього дослідження. Його концепція представляє собою застосування латентно-семантичного аналізу та послідовності просторових операцій на двомірній площині для отримання семантичної структури наукового тексту, приклад якої зображено на рис.2.

Результуюча семантична мережа складеться із кластерів-слів (позначено на малюнку як WordCluster) до яких прив'язана вага кластера (зображена на малюнку у квадратних дужках), кластерів речень (позначено на малюнку як SentenceCluster) які пов'язані із кластерами-стемами семантичним відношенням, вага якого позначена у круглих дужках.

Отримана семантична мережа використовується для зняття числових даних, що характеризують семантичні властивості документа, до яких відносяться кількість стем-кластерів та кількість стем-речень, ваги кластерів-стем, ваги і кількість зв'язків, кількість стем що не увійшли до семантичної мережі, які використовуються для автоматизації процесів оцінки і фільтрації текстів за їх семантичною складовою Головною властивістю створеної структури є те, що вона дозволяє встановити прямі відносини між кластерами-стемами і масивами речень, що і являє собою семантичне подання документу, без застосування попередньої семантичної розмітки або залучення попередніх семантичних знань.

Крок третій – для кожного кластера-стеми із семантичних моделей текстів проводиться перевірка входження його елементів в користувальницький запит, для чого використовується розроблений механізм визначення загальної частини на основі відстані Левенштейна [6], що застосовується для кожного ключового слова і кожної стеми у кластері. Якщо для поточного кластера-стеми таке входження знайдено - то

10.32782/KNTU2618-0340/2020.3.2-1.13

кластер стає кандидатом для включення у результуючу відповідь. Приклад кластеру-стеми тексту на тематику «чорні діри» зображено у таблиці 1.

Таблиця 1

Приклад фрагменту кластеру-стеми

диску среднемасивних стала аналог важливі витягнутої всередині тертя дозволило Зовнішнє великих завершила стала вік тертя обсерваторія Через що з'єднують виявити з'єднують витягнутої забезпечує даними пророкує будова завершила забезпечує показана среднемасивних вік Альберта об'єкти назад тертя будова завершила рентгенівському речовини Метагалактики Сонця утворення подібних поглинає будова забезпечує Чумацького відстанях забезпечує область тертя астрономів білої будова тертя вік Альберта досліджень обсерваторія тертя обсерваторія тертя пророкує Альберта завершила нашої темпом багатьох спостерігача забезпечує становить інших сусідній стала компаньйоном Чорні вік Альберта масивних тертя десятків обсерваторія Виявлення тертя будова існування більш можливо відкрито скупчення визначаються завершила аналог високим тепер джерел

Як можна побачити з наведеного прикладу, у одному кластері містяться стеми, що мають прямий семантичний зв’язок, проте синтаксично вони ніяк не співвідносяться (чорні, обсерваторія, метагалактики, Сонця тощо). Саме це дає можливість системі знаходити семантично близькі речення у тексті, не спираючись при цьому на синтаксичні представлення (якщо ми шукаємо чорні діри – то інформацію про метагалактики також необхідно вважати релевантною, при умові достатньої сили семантичних зв’язків). Інші «нерелевантні» словоформи (наприклад – розповсюджені дієслова) відсікаються за недостатньою вагою відносно тексту в цілому і конкретного запиту. У цьому і полягає реалізація моделі м’якого розуміння Леонтьєвої, коли ситуація – а власне, запит користувача, змінює сенсові ваги одного і того самого тексту.

Важливо зазначити, що функціонування системи стає можливим, оскільки до її складу була включена база знань, що представляє собою корпус документів, тексти з якої мають відповідно побудовану семантичну модель документа. Корпус складається з декількох класів текстів, обробка яких по-перше повно покриває стилістичні і семантичні особливості мови, а по-друге стане опорою для проведення повноцінного тестування системи. Для цього була складена колекція із 100 текстів наукового стилю, розміри яких розподілені від 6 до 203 Кілобайт чистого тексту, класи яких розподілені у відсотковому співвідношенні у таблиці 2:

Таблиця 2

Відсоткове розподілення типів тексту у колекції

Тип	Кількість (%)
0	35
1	15
2	21
3	24
4	5

Де тип 0 – відповідає тексам із слабкою семантичною зв’язністю, що були створені за допомогою автоматичного генерування або складені із фрагментів різних текстів і використовуються для тестування системи, тип 1 – тексти на економічну тематику, тип 2 – тексти на тематику філософії, тип 3 – тести на тематику космології і астрономії, а тип 4 – тексти на тематику інформаційних технологій і програмування. Таке розподілення переслідує мету створення достатньо репрезентативної вибірки текстів та розробки необхідного набору даних, що дозволить створити гнучку систему оцінок та перевірок адекватності додатку. Велика кількість семантично помилкових текстів створює умови для виявлення випадковостей у створенні відповідей, а велика кількість тематично незалежних кластерів тексту дозволяє змоделювати можливості обробки запиту користувача у полістилистичній колекції текстів. Окрім того, у ході дослідження була перевірена адекватність побудови семантичної моделі документа, що

10.32782/KNTU2618-0340/2020.3.2-1.13

викладена у роботі [7] та заснована на оцінці сенсової місткості семантичних міток документу. Для цього, кожному документу із корпусу була побудована відповідна семантична модель тексту. Із кожної побудованої таким чином мережі були отримані кластери-стеми із найбільшою (семантично сильний кластер) та найменшою (семантично слабкий кластер) загальною вагою перетинів із семантичними контурами стем, для яких було розраховане значення сенсової місткості S_V за формулою (1)

$$S_V = \frac{N_Q}{N_W}, \tag{1}$$

де N_W – загальна кількість слів у документі, необхідна для нормалізації отриманих результатів, N_Q- встановлена емпірично кількість унікальних термінів у кластері, що мають пряме відношення до галузі знань, до якої належить текст (тематика документа). Отримані результати розрахунку сенсової місткості для кожного тематичного набору текстів зображені на рис.3.

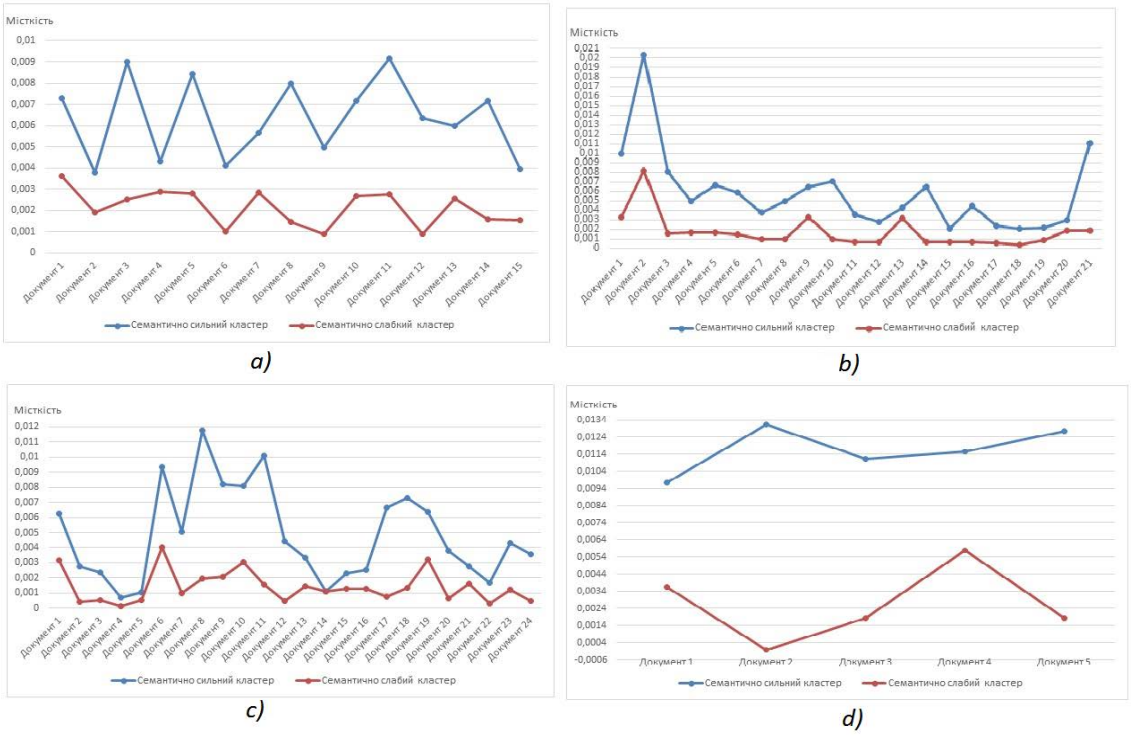

Рис. 3. Розподіл значень сенсової місткості за темами: a) економіка; b) філософія c) астрономія; d) інформаційні технології.

Отримані результати показали, що у проведених тестах кількість семантично значимих 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 3,375 рази, для тематики «астрономія» (24 документа) – у 3,473 рази, для тематики «інформаційні технології» (6 документів) - у 4,44 рази. При проведенні досліджень не враховувалась вага стем або їх будь-яке синтаксичне співвідношення із текстом – пов'язані терміни оцінювалися лише з точки зору належності до тематики документу, тому отримані результати вказують на цілком

10.32782/KNTU2618-0340/2020.3.2-1.13

адекватне формування семантичних міток документу – кількість семантично значимих термінів у кластері-стемі прямо пропорційна до кількості та ваги пов’язаних із ним семантичних контурів речень у побудованих моделях документів, що доводить залежність структури семантичної мережі саме від семантичної складової тексту.

Крок четвертий – для подальшого аналізу система повинна ранжувати кластери-кандидати за їх співвідношенням із вхідним запитом, тому для кожного знайденого кластера-стеми розраховується ситуативна вага W_S – кількісна міра, що характеризує загальну вагу стем у тексті що збіглися із запитом користувача, після чого усі знайдені на попередньому кроці кластери-стеми сортуються відповідно до розрахованого значення ситуативної ваги. Для обмеження і зручності аналізу відповідей розмір результуючого тексту, що буде згенеровано для користувача, виходить із кількості кластерів із максимальною ситуативною вагою C_W, що вираховується за формулою (2):

$$C_W = \frac{W_C}{S_C}\ , \tag{2}$$

де W_C – загальна кількість стем у базі системи, S_C – загальна кількість речень у базі системи.

Заключний крок – генерація текстової відповіді користувачеві. Для цього вибираються кластери-речення, пов'язані з кожним кластером-стемою загальною кількістю C_W, що стали кандидатами для включення в результуючу відповідь на попередньому етапі. Якщо таких пов'язаних кластерів кілька, то в множину кандидатів потрапляє кластер з максимальною вагою зв’язку. В результаті даної операції формується множина речень-кандидатів для включення у результуючу відповідь, що сортується за номером речення у вихідному документі і загальною вагою стем у реченні. На виході користувач отримує згенеровану текстову відповідь, семантично пов’язану з вхідним запитом. Приклад отриманої відповіді на запит «інтерфейс» наведено у табл. 3.

Таблиця 3

Приклад відповіді системи на запит «інтерфейс»

Інші абстрактні типи даних - метакласи, інтерфейси, структури, перерахування, - характеризуються якимись своїми, іншими особливостями. Поряд з поняттям «об'єкта» клас є ключовим поняттям в ООП (хоча існують і безкласового об'єктно-орієнтовані мови, наприклад, Self, Lua; докладніше дивіться прототипна програмування). Суть відмінності класів від інших абстрактних типів даних полягає в тому, що при завданні типу даних клас визначає одночасно як інтерфейс, так і реалізацію для всіх своїх екземплярів, а виклик методу-конструктора обов'язковий. На практиці об'єктно-орієнтоване програмування зводиться до створення певної кількості класів, включаючи інтерфейс і реалізацію, і подальшого їх використання. Використовувані людиною класифікації в зоології, ботаніки, хімії, деталях машин, несуть в собі основну ідею, що будь-яку річ завжди можна уявити окремим випадком деякого більш загального поняття. Саме тому приклади класів в навчальних посібниках з об'єктно-орієнтованого програмування так часто згадують яблука і груші. В об'єктно-орієнтованої програми із застосуванням класів кожен об'єкт є «екземпляром» деякого конкретного класу, і інших об'єктів не передбачено. При цьому в різних мовах програмування допускається або не допускається існування ще якихось типів даних, екземпляри яких не є об'єктами (тобто мова визначає, чи є об'єктами такі речі, як числа, масиви і покажчики, або не є, і, відповідно, чи є такі класи як «число», «масив» або «покажчик», екземплярами яких були б кожне конкретне число, масив або покажчик). При використанні класів все елементи коду програми, такі як змінні, константи, методи, процедури і функції, можуть належати (а в багатьох мовах повинні належати) того чи іншого класу. Метод, співвіднесений з екземпляром класу (звичайний метод), може бути викликаний тільки у самого об'єкта, і має доступ як до статичних полів класу, так і до звичайних полях конкретного об'єкта (при виклику цей об'єкт передасться прихованим параметром методу). Успадкованих клас або інтерфейс буде містити в собі все, що зазначено для всіх його батьківських класів (в залежності від мови програмування і платформи, їх може бути від нуля до нескінченності). Часто різні змінні програми зберігають логічно пов'язані значення, і за підтримку цієї логічної зв'язності несе відповідальність програміст, тобто автоматично зв'язність не підтримує. Клас - це елемент ПО, що описує абстрактний тип даних і його часткову або повну реалізацію. Точний зміст цієї фрази буде розкритий нижче. Графічне представлення деякої кількості класів та зв'язків між ними називається діаграмою класів. Ідея класів прийшла з робіт по базам знань, що мають відношення до досліджень з штучного інтелекту. Тобто «екземпляр класу» в даному випадку означає не «приклад деякого класу» або «окремо взятий клас», а «об'єкт, типом якого є якийсь клас». Наприклад, абстрактний тип даних «рядок тексту» може бути оформлений у вигляді класу, і тоді все рядки тексту в програмі будуть об'єктами - екземплярами класу «рядок тексту». Сам клас в підсумку визначається як список своїх членів, а саме полів (властивостей) і методів / функцій / процедур. Наприклад, загальна кількість

10.32782/KNTU2618-0340/2020.3.2-1.13

рядків тексту, створених в програмі за час її роботи, буде статичним полем класу «рядок тексту». В ООП при використанні класів весь виконуваний код програми (алгоритми) буде оформлятися у вигляді так званих «методів», «функцій» або «процедур», що відповідає звичайному структурному програмуванню, однак тепер вони можуть (а в багатьох мовах зобов'язані) належати тому чи іншого класу. Як і поля, код у вигляді методів / функцій / процедур, що належать класу, може бути віднесений або до самого класу, або до екземплярів класу. У програмуванні існує поняття програмного інтерфейсу, що означає перелік можливих обчислень, які може виконати та чи інша частина програми, включаючи опис того, які аргументи і в якому порядку потрібно передавати на вхід алгоритмам з цього переліку, а також що і в якому вигляді вони будуть повертати. Абстрактний тип даних інтерфейс придуманий для формалізованого опису такого переліку. Програмні інтерфейси, а також класи, можуть розширюватися шляхом успадкування, яке є одним з важливих засобів повторного використання готового коду в ООП. В об'єктно-орієнтованої програми прапорець «звільнений» буде оголошений приватним членом деякого класу, а для читання і зміни його будуть написані відповідні публічні методи. Скрізь далі слова «клас», «об'єкт», «інтерфейс» і «структура» будуть вживатися в своїх спеціальних значеннях, заданих в рамках ООП

Отримана відповідь містить інформацію не тільки дані про інтерфейс що цікавить користувача, а й семантично пов’язану інформацію - щодо абстракцій у теорії програмування, об’єктно-орієнтованого підходу до розробки, інформацію про класи, тощо. Згенерований текст був отриманий на основі документів «Клас» і «Інтерфейс» із загального корпусу документів. Окрім того, алгоритм є стійким до кількості слів у запиті і їх семантичних зв’язків – у табл. 4 наведено приклад запиту що є комбінацією тематично однозначного (космос) і неоднозначного (дослідження) ключових слів.

Таблиця 4

Приклад відповіді системи на запит «дослідження космосу»

І нарешті, як і роль самої космонавтики для людства. Основним поштовхом до цього було протистояння двох наддержав (СРСР і США) - холодна війна. Початковий період обертання супутника навколо Землі склав 96,2 хв, а нахил - 65'1 '. «Пальма першості» в області освоєння космосу дісталася СРСР, але США теж не хотіли відставати, і в такий подією світової важливості став політ на Місяць. Таким чином, «Луна-3» виявилася першим апаратом, який став штучним супутником відразу і для Землі, і для Місяця - його сильно витягнута орбіта охоплювала обидва цих небесних тіла. Так люди вперше змогли побачити зворотний бік Місяця. «Родзинкою» її було дзеркальце, погойдуються вгору-вниз і при цьому повільно що повертається навколо вертикальної осі зліва направо. З його допомогою створювалася порядкова запис всього зображення. Англійські астрономи, зробивши їх обробку і отримавши зображення місячної поверхні, не стали передавати його в друк, чекаючи, коли першими ці сенсаційні дані опублікують російські. Відповіді не було, тому, вважаючи себе вільними від подальшого дотримання коректності, британські дослідники передали знімок в газети. В результаті все зображення вийшло стислим з боків і розтягнутим у висоту - місячна поверхня постала у вигляді стирчать вгору вузьких загострених каменів, між якими були ще більш вузькі піщані обеліски

Брежнєв, але поки тривало узгодження, настав пізній вечір, і його вирішено було не турбувати. Багато тисяч років тому, коли бачиш нічне небо, людина мріяв про політ до зірок. У 19 столітті з'явився фантастичне оповідання письменника Жюля Верна "З гармати на Місяць". Однак ці ракети були технічно необгрунтованої мрією. Вчені за багато століть не назвали єдиного розташованого у розпорядженні людини засобу, за допомогою якого можна подолати могутню силу земного тяжіння і полинути в міжпланетний простір. У 1911 році Ціолковський вимовив свої віщі слова: "Людство не залишиться вічно на Землі, але, в гонитві за світлом і простором, з початку боязко проникнути за межі атмосфери, а потім завоює собі всі близько земне простір". І з цього моменту великі уми планети почали працювати над початком реального освоєння космосу. Близько 20 польотів до Місяця американських автоматичних станцій за програмами «Рейнджер», «Сервейер» і «Лунар Орбитер» були строго підпорядковані підготовці до висадки людини на Місяць. Доставити туди експедицію повинна була гігантська ракета «Сатурн-V», створена під керівництвом Вернера фон Брауна, німецького конструктора снарядів «Фау», який після другої світової війни працював в США. Втім, Радянський Союз також не стояв осторонь від підготовки пілотованого «місячного» польоту

І тут астронавти в порушення програми попросили дозволу почати вихід на Місяць негайно. У романі «Із Землі на Місяць» французький письменник Жюль Верн так описав перший політ людей навколо нашого супутника. Перший пілотований корабель, облетів навколо Місяця, був запущений в Сполучених Штатах - як у романі, так і в дійсності. Фраза, за словами самого Армстронга, була «добре підготовленим експромтом», заздалегідь обраним із сотень надійшли до польоту пропозицій. Такий зразок називався аварійним і повинен був братися, не відходячи від місячного модуля на випадок, якщо якісь надзвичайні обставини змусять астронавтів терміново сховатися всередині кабіни і покинути Місяць (такі ж зразки згодом бралися і всіма іншими п'ятьма «Аполлонами»). Гармата, з якої був випущений «місячний» снаряд, називалася Колумбіада, командний модуль корабля «Аполлон-11» носив ім'я «Колумбія». Снаряд у Жюля Верна приводнився 11 грудня але був підібраний кораблем лише 29 грудня, що майже збігається з датою приводнення «Аполлона-8» - 27 січень. Повертаючись на Землю, як фантастичний, так і реальні космічні кораблі здійснювали посадку на воду в північній половині Тихого океану.

Результуючий текст був отриманий на основі документів «Досягнення у освоєнні космосу» і «Астрономічна картина миру» та містить інформацію про космічну гонку СРСР та США, висадку людини на місяць, відсилки до романів Жуля Верна, та інші релевантні дані, що вказує на коректний процес роботи системи запит-відповідь.

Для перевірки отриманої системи були проведені індивідуальні оцінки відповідей системи за бальним методом, що 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, а саме: присутність відповіді,

10.32782/KNTU2618-0340/2020.3.2-1.13

ступень збігу із тематикою запиту, повнота викладу, присутність тематичних розривів та присутність сенсових розривів. Усього у ході дослідження було проведено 100 тестів із різною складністю питань і різним тематичним напрямом запитів, на кожен з яких була отримана відповідна оцінка, фрагмент графіку значень якої зображено на рис. 4.

Рис. 4. Фрагмент графіку значень експертних оцінок системи для 50 тестів.

Середнє значення усіх проведених експертних оцінок склало 0,839, що вказує на задовільні результати роботи системи. Проведення мануального тестування показало наявність проблем у присутності сенсових розривів та необхідності обробки виключних ситуацій із відсутністю необхідних знань у системі, що і знизило сумарно оцінку роботи системи, проте ці факти не мають критичного впливу на алгоритм генерації відповідей. Виконання тестів системи запит-відповідь із використанням автоматичної генерації текстів показали адекватність роботи додатку як з точки зору використання семантичної моделі як основи для побудови інтелектуальних пошукових інструментів, так і з точки зору доцільності застосування семантичних моделей у ситуативній генерації текстів. Більш детально процес тестування створеної системи описаний у роботі [8].

Питання створення системи, що базується на деяких знаннях йде пліч о пліч із питанням адаптивності шляху накоплення цих знань. Розробка підходу, заснованого на побудові семантичної моделі документа створює великі можливості у цьому плані, дозволяючи незалежно наповнювати колекцію документів (і базу знань системи відповідно) текстами наукового стилю мовлення різної тематичної спрямованості. Система функціонує таким чином, що у користувача немає необхідності у ручній семантичній розмітці тексту, класифікації або перевірки тексту на відповідність якимось критеріям – достатньо лише завантажити документ у форматі plain text. Проте, такий підхід призводить до можливості наповнення системи знаннями, що недостатньо якісні для її стабільної роботи, що може створити надлишкову множину текстів для аналізу. До таких текстів, в першу чергу, відносяться недостатньо формалізовані документи, відповідність викладення інформації в яких суперечить науковому стилю

10.32782/KNTU2618-0340/2020.3.2-1.13

мовлення. Вирішенням цієї проблеми стало застосування процесу попередньої оцінки і автоматичної фільтрації документів за їх семантичною зв'язністю.

За створеним алгоритмом, кожен з текстів характеризувався двома значеннями – нормалізованим розміром тексту W_N, отриманого за формулою (3):

$$W_N = \frac{W_i - W_{\min}}{W_{\max} - W_{\min}}, \quad (3)$$

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 (4):

$$S_N = \frac{W_U - CW_C}{W - CW}, \quad (4)$$

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

Отримані таким чином данні є критеріями оцінки науковості тексту і водночас - навчальною вибіркою для включеної у структуру системи нейронної мережі, що має один прихований рівень, кількість нейронів у якому дорівнює кількості уроків що подано на вхід, та один вихід що може видавати значення 0 або 1 – наші класи тексту відповідно. Для навчання мережі використовується друге правило Хебба (дельта-правило), функцією активації є сигмоїдна функція, а значення зміщення одноразово обираєтеся випадково у проміжку між –0.5 та 0.5. Отримана навчена модель використовується для прогнозування зв'язності вхідних текстів, у випадку, коли користувач бажає доповнити базу знань системи – якщо відповідно до отриманого прогнозу документ що завантажується є достатньо формалізованим, він проходить побудову семантичної моделі та завантажується у базу знань системи, інакше користувач отримує помилку і завантаження не відбувається. Необхідно враховувати, що описані семантичні характеристики залежать від розміру тексту, тому зняті данні потребують попередньої нормалізації. В якості даних для навчання використовувалися тексти із описаного раніше корпусу – 50 текстів було використано як набір уроків (40% склали незв'язні тексти, а 60% - зв'язні). Для тестування отриманої мережі було виконано 100 тестів, данні для яких вибиралися з бази даних колекції текстів. Проведені тести показали точність роботи у 90 %, причому жоден прогноз щодо незв'язних текстів не був хибним – це стосується як цілком автоматично згенерованих текстів, що вказує на коректну обробку статичних властивостей тексту, так і фрагментарних текстів, що говорить про коректну обробку семантично незв'язних текстів. Такі результати вказують не тільки на достатню і задовільну точність роботи класифікатора і можливість його подальшого використання у системі, а і на адекватність семантичної моделі документу в цілому і її доцільності у застосуванні в інтелектуальних системах АОТ. Більш детально застосування системи класифікації наукових текстів розглянуто у [9–10].

Висновки

У роботі представлено комп'ютерну модель системи «запит-відповідь» із використанням автоматичної генерації текстів. Для вирішення головних наукових проблем моделі була розроблена та реалізована концепція безсловникової методики побудови семантичної моделі документу, що не має аналогів і використовує у своїй

10.32782/KNTU2618-0340/2020.3.2-1.13

роботі як базові методи синтаксичної обробки тексту, так і науково нові застосування латентно-семантичного аналізу, що у комплексі дозволили створити семантичну модель документу без попередньої розмітки та залучення додаткових словникових знань. Аналіз схожих розробок семантичних мереж вказує на інноваційність підходу, а проведені тестування показали, що алгоритм стійкий до семантичних змін і дійсно залежить саме від сенсу тексту, а не від випадковості або частотного подання. Окрім того, реалізована модель представлена у вигляді відкритого програмного API [https://github.com/yegorkovylin/sm-scr.git], що дозволить розробникам систем АОТ отримувати кількісні моделі семантичних характеристик тексту без необхідності залучення лінгвістичних знань – прикладом такого застосування може служити розроблений у цій роботі засіб автоматичної оцінки семантичної зв'язності тексту, що використовує нейрону мережу, навчену на основі числових даних отриманих із семантичної моделі документу. Проведені тести роботи вказують на надійність процесу класифікації і доводять перспективність досліджень використання створеного алгоритму побудови семантичної моделі тексту для вирішення інших завдань АОТ, не розглянутих в рамках цієї роботи. Проведене дослідження розробки системи запит-відповідь на основі генерації текстів показує доцільність використання процесу породження текстів для формування відповіді на поставлене питання. У ході виконання дисертації був розроблений алгоритм і програмний додаток, що дозволяє створювати, на основі семантичних моделей нерозмічених колекцій документів, нові тексти, що містять відповіді на поставлені користувачем питання. Отримані результати проведеного тестування вказують на коректність і адекватність отриманого додатку.

Список літератури

1. Поляков П. Ю. Використання семантичних категорій в завданні класифікації відгуків про книги. *Матеріали міжнародної конференції «Діалог»* (м. Москва, 29 травня – 2 червня 2013 р.). Москва, 2013. С. 193–199.
2. Антонов А. В. Галактика Zoom. Оцінка модифікації методу формування інфопортрета. *Матеріали третього російського семінару по оцінці методів інформаційного пошуку*. (м. Ярославль, 6 жовтня 2018 р.). Ярославль, 2018. С. 226.
3. Губин М.Ю., Разин В.В., Тузовский А.Ф. Применение семантических сетей и частотных характеристик текстов на естественных языках для создания семантических метаописаний. *Проблемы информатики*. 2011. № S2. С. 59–63.
4. Pismak A. E., Kharitonova A. E. The Method of Automatic Formation of a Semantic Network from Weakly Structured Sources. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*. 2016. Vol. 16. № 2. P. 324–330.
5. Волковський О. С., Ковилін Є. Р. Комп'ютерна система інтелектуального семантичного пошуку з використанням генерації текстів. *Вісник Херсонського національного університету*. 2018. № 3(66). С. 238–245.
6. Volkovsky O. S., Kovylin Y. R. Computer System of Building of the Semantic Model of the Document. *IEEE Second International Conference on Data Stream Mining & Processing*. (Lviv, August 21-25, 2018). P. 322–327. DOI: 10.1109/DSMP.2018.8478591.
7. Volkovsky O. S., Kovylin Y. R. Mathematical Model for Automatic Creation the Semantic Thesaurus for the Scientific Text. *System Technologies*. 2019. № 6. P. 82–88.
8. Волковський О. С., Ковилін Є. Р. Модель автоматичної оцінки адекватності комп'ютерних систем «запит-відповідь» з використанням генерації текстів. *Системні технології*. 2020. № 4 (129). С. 50–58.
9. Волковський О. С., Ковилін Є. Р. (2017). Комп'ютерна система автоматичного визначення зв'язності тексту. *Системні технології*. 2017. № 1 (112). С. 11–17.

10.32782/KNTU2618-0340/2020.3.2-1.13

10. Волковський О. С., Ковилін Є. Р. (2018). Комп'ютерна система автоматичного аналізу промислових інструкцій. *Системні технології.* 2018. № 3(116). С. 28–37.

References

1. Poliakov, P. Iu. (2013). Vykorystannia semantychnykh katehorii v zavdanni klasyfikatsii vidhukiv pro knyhy. Proceedings of the *international conference «Dialoh»*. (Moscow, May 29 – June 2, 2013), pp. 193–199.
2. Antonov, A. V. (2018). Otsinka modyfikatsii metodu formuvannia infoportreta: Halaktyka Zoom. Proceedings of the *third Russian seminar on the evaluation of information retrieval methods*, (Yaroslavl, October 6, 2018), pp. 226.
3. Gubin, M. Yu., Razin, V. V., & Tuzovsky, A. F. (2017). Application of semantic networks and frequency characteristics of texts on natural languages for the creation of semantic metapopsis. *Problems of Informatics*. **S2**, 59–64.
4. Pismak, A. E., & Kharitonova, A. E. (2016). The method of automatic formation of a semantic network from weakly structured sources. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics.* **16**, 2, 324–330.
5. Volkovskyi, O. S., & Kovylin, Ye. R. (2018). Kompiuterna systema intelektualnoho semantychnoho poshuku z vykorystanniam heneratsii tekstiv. *Visnyk Khersonskoho natsionalnoho universytetu*. **66**, 3, 238–245.
6. Volkovsky, O. S., & Kovylin, Y. R. (2018). Computer System of Building of the Semantic Model of the Document. *2018 IEEE Second International Conference on Data Stream Mining & Processing* (Lviv, August 21-25, 2018), pp. 322–327. DOI: 10.1109/DSMP.2018.8478591
7. Volkovsky, O. S., & Kovylin, Y. R. (2019). Mathematical model for automatic creation the semantic thesaurus for the scientific text. *System technologies*. **6**, 82–88.
8. Volkovskyi, O. S., & Kovylin, Ye. R. (2020). Model avtomatychnoi otsinky adekvatnosti kompiuternykh system «zapyt-vidpovid» z vykorystanniam heneratsii tekstiv. *Systemni tekhnolohii.* **4**, 50–58.
9. Volkovskyi, O. S., & Kovylin, Ye. R. (2017). Kompiuterna systema avtomatychnoho vyznachennia zviaznosti tekstu. *Systemni tekhnolohii.* **1**, 11–17.
10. Volkovskyi, O. S., & Kovylin, Ye. R. (2018). Kompiuterna systema avtomatychnoho analizu promyslovykh instruktsii. *Systemni tekhnolohii.* **3**, 28–37.

Ковилін Єгор Романович – аспірант, Дніпровський національний університет ім. Олеся Гончара, Україна, kovilin.yegor@gmail.com.

Волковський Олег Степанович – к.т.н., доцент, доцент кафедри комп'ютерних наук та інформаційних технологій Дніпровський національний університет ім. Олеся Гончара, Україна, ffeks@365.dnu.edu.ua.

10.32782/KNTU2618-0340/2020.3.2-1.13