МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів аналізу й обробки експертної лінгвістичної інформації

(тема)

Виконав:
Студент 2 курсу, групи ІПЗм-20-4
Євсютін І.Д.
(прізвище, ініціали)

Спеціальність 121 Інженерія програмного забезпечення
(код і повна назва спеціальності)

Тип програми освітньо-наукова

Керівник проф. Шостак І.В.
(посада, прізвище)

Допускається до захисту

Зав. кафедри _____ З.В. Дудар
(підпис) (прізвище, ініціали)

2022 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Програмної інженерії
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 121 Інженерія програмного забезпечення
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Інженерія програмного забезпечення
(повна назва)

ЗАТВЕРДЖУЮ:
Зав.кафедри _____
(підпис)
« ____ » _____ 2022 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студента Євсютіну Івану Дмитровичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів аналізу й обробки експертної лінгвістичної інформації

затверджена наказом університету від 24.03.2022 р. № 412 Ст

2. Термін подання роботи до екзаменаційної комісії 10 05 2022р.
3. Вихідні дані до роботи *теорія лінгвістики, теорія автоматів, онтології*
4. Перелік питань, що потрібно опрацювати в роботі *мета роботи, аналіз проблемної галузі і постановка задачі, опис запропонованих варіантів оптимізації, використовувані методи та алгоритми, опис розробленої програмної системи, опис застосованих програмних рішень, аналіз можливих застосувань*

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі	31 березня 2022	*виконано*
2.	Огляд існуючих методів	10 квітня 2022.	*виконано*
3.	Розробка алгоритмів, проектування та розробка ПЗ	15 квітня 2022	*виконано*
4.	Підготовка пояснювальної записки	20 квітня 2022	*виконано*
5.	Спецчастина	28 квітня 2022	*виконано*
6.	Підготовка презентації та доповіді	03 травня 2022	*виконано*
7.	Попередній захист	05 травня 2022	*виконано*
8.	Нормоконтроль, рецензування	07 травня 2022.	*виконано*
9.	Занесення роботи в електронний архів	08 травня 2022.	*виконано*
10.	Допуск до захисту в зав. кафедри	11 травня 2022	*виконано*

Дата видачі завдання 28 березня 2022р.

Студент _____
(підпис)

Керівник роботи _____ проф. Шостак І.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Кваліфікаційна робота магістра містить: 103 с., 24 рис., 5 табл., 35 джерел.

ОНТОЛОГІЇ, КЛАСИФІКАЦІЯ, НЕЙРОННА МЕРЕЖА, ЛІНГВІСТИЧНИЙ АНАЛІЗ.

Об'єктом дослідження є методи онтологічних описів.

Метою роботи є дослідження й розробка математичних методів і алгоритмів аналізу лінгвістичної експертної інформації, отриманої з колекцій текстів заданої предметної області, з метою її формалізації для застосування в інтелектуальних системах обробки інформації.

Методи розробки – методи імовірнісного тематичного моделювання, методи кластерного аналізу, статистичний експеримент.

У результаті реалізовано алгоритм аналізу й обробки лінгвістичної експертної інформації стосовно до завдання побудови простої онтології

ONTOLOGIES, CLASSIFICATION, NEURAL NETWORK, LINGUISTIC ANALYSIS.

The object of research is the methods of ontological descriptions.

The aim of the work is to study and develop mathematical methods and algorithms for the analysis of linguistic expert information obtained from collections of texts of a given subject area, in order to formalize it for use in intelligent information processing systems.

Methods of development - methods of probabilistic thematic modeling, methods of cluster analysis, statistical experiment.

As a result, the algorithm of analysis and processing of linguistic expert information in relation to the task of constructing a simple ontology is implemented.

Умови публікації пояснювальної записки

Я, _____Євсютін Іван Дмитрович_____
(прізвище, ім'я, по батькові)

студент групи ___ІПЗм-20-4_здобувач вищої освіти на другому (магістерському) рівні

кафедра програмної інженерії,
(повна назва кафедри)

заявляю: моя кваліфікаційна робота на тему ___ Дослідження методів аналізу й обробки експертної лінгвістичної інформації ___,
(назва роботи)

що буде представлена до ЕК для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

ЛЕІ – лінгвістична експертна інформація

ПМ – природня мова

ЛСА – латентно-семантичний аналіз

ВТМ – ймовірносно-тематичне моделювання

ВЛСА – імовірнісний латентно-семантичний аналіз

ЛРД – латентне розміщення Дирихле

SVD – сингулярне розкладання (singular value decomposition)

МРЧ – метод рою часток

ПрО – предметна область

ІСП – ієрархічна структура понять

ІАК – ієрархічна агломеративна кластеризація

ВСТУП

У цей час повсюдне застосування цифрових технологій у різних областях діяльності людину привело до росту обсягу інформації, що накопичується, і даний обсяг інформації постійно збільшується. Згідно з дослідженнями міжнародної дослідницької компанії International Data Corporation (IDC), що займається вивченням світового ринку інформаційних технологій і телекомунікацій, обсяг інформації в мережі Інтернет подвоюється кожні вісімнадцять місяців. Найголовніше, далеко не всі дані цінні – по оцінках IDC, до 2020 року частка корисної інформації складе всього 35% від усієї згенерованой. Як наслідок, керування різними сферами діяльності людини стає неможливим без процесу аналізу й обробки інформації. Ріст кількості джерел інформації (соціальні мережі, платформи мікроблогів та ін.), збільшення її обсягів, наявність різнорідних даних (структуровані, неструктуровані й квазіструктуровані) приводить до необхідності поліпшення існуючих методів її обробки, а також розробки нових, відповідних до сучасних потреб, інтелектуальних методів аналізу інформації [1].

Інтелектуальна обробка текстової інформації, яка знаходить своє застосування практично у всіх сферах інформаційно-аналітичної діяльності людини (наукові дослідження, сфера освіти, охорона здоров'я, ділова галузь та багато інші), охоплює широке коло проблем, таких як: формалізація текстових даних (наприклад, у вигляді онтологій, семантичних мереж, тезаурусів та ін.), способи витягу знань із даних (наприклад, кластеризація і класифікація, отримання фактографічної інформації, асоціативні правила та ін.), методи інтеграції предметних знань, отриманих з різнорідних джерел (наприклад, відображення, злиття, вирівнювання).

Одним з перспективних рішень названої проблеми є застосування технологій, заснованих на онтологіях [2]. Формалізоване представлення великого обсягу текстів у вигляді онтологічної моделі дозволяє виконувати автоматичну

(інтелектуальну, аналітичну) обробку отриманої онтологічної інформації, що знаходить застосування в інформаційно-аналітичних і пошукових системах нового покоління й системах інтеграції даних. Досягнуті вагомі результати в області автоматичного витягу структурованих даних і онтологічного представлення знань,

Незважаючи на високу наукову й практичну значимість існуючих робіт у даній області, у цей час відсутній універсальний для всіх предметних областей метод формалізації витягнутих з текстової інформації знань, що дозволяє з високою точністю й повнотою витягати терміни й зв'язку між ними, що й передбачає можливість розширення формалізованих структур різних джерел шляхом їхньої інтеграції. Із цієї причини існуючі методи не можуть забезпечити достатньої швидкості обробки більших обсягів даних. Виникає ситуація, коли користувач має доступ до інформації, але не може вчасно скористатися її корисною складовою.

Метою роботи є дослідження й розробка математичних методів і алгоритмів аналізу лінгвістичної експертної інформації, отриманої з колекцій текстів заданої предметної області, з метою її формалізації для застосування в інтелектуальних системах обробки інформації.

Завдання, розв'язувані для досягнення поставленої мети.

Виконати аналіз існуючих методів і проблем в області аналізу, обробки й формалізації лінгвістичної експертної інформації.

Розробити аналітичний метод формалізації лінгвістичної експертної інформації, отриманої з колекції текстів заданої предметної області, у вигляді ієрархічної структури понять.

Розробити алгоритм інтелектуальної обробки лінгвістичної експертної інформації для формалізації колекцій текстів заданої предметної області.

Реалізувати запропоновані методи й алгоритм аналізу й обробки лінгвістичної експертної інформації стосовно до завдання побудови простої онтології й оцінити його ефективність.

Методами дослідження є: системний аналіз, методи математичної статистики й теорії ймовірностей, машинне навчання, методи імовірнісного тематичного моделювання, методи кластерного аналізу, статистичний експеримент, аналіз і узагальнення.

Програмний модуль, що має бути розроблений на основі запропонованого алгоритму обробки ЛЕІ, може використовуватися як для рішення завдань інтелектуальної обробки текстової інформації (витяг тематичних кластерів і тем з колекцій текстів, кластерізація, візуалізація змісту колекції текстів, інтеграція ієрархічних структур понять), так і для побудови простої онтології колекції текстів, заданої у вигляді ієрархічної структури понять предметної області.

В результаті має бути запропонований алгоритм формалізації лінгвістичної експертної інформації, отриманої з колекції текстів заданої предметної області, у вигляді ієрархічної структури понять, що відрізняється від відомих обчисленням схованих закономірностей у тексті, кластерізацією понять і відображенням їх ієрархічних зв'язків у вигляді дендрограми з обліком тематичної імовірнісної близькості.

1 АНАЛІЗ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОЇ ОБРОБКИ ЛІНГВІСТИЧНОЇ ІНФОРМАЦІЇ

1.1 Завдання інтелектуального аналізу лінгвістичної інформації

Завдання інтелектуальної обробки лінгвістичної експертної інформації (ЛЕІ), пов'язана з обробкою текстової інформації й витягом знань їх текстів, є досить складною. Кожний етап усього технологічного циклу обробки, починаючи з морфологічного розбору слова, і закінчуючи співвіднесенням терміна із семантичною категорією, сполучений з поруч методологічних і технологічних проблем. Складність завдання витягу інформації з текстів залежить від різних факторів, наприклад, складності і якості тексту, особливостей мови, на якій підготовлений текст документа, типу інформації, що витягає, видів зв'язків між сутностями й т.п. Неоднозначність виявлення й інтерпретації слова й тексту в цілому є дуже серйозною проблемою, без дозволу якої досягнення значимих успіхів в області аналізу текстової інформації малоймовірно [3].

У загальному випадку процес інтелектуального аналізу текстової інформації можна представити як послідовність декількох кроків: попередня обробка, формалізація неструктурованої інформації та інтеграція формалізованих представлень. Етапи попередньої обробки текстової інформації й одержання ЛЕІ з текстів, показано на рис. 1.1.

У ході виконання попередньої обробки одержується ЛЕІ. Під лінгвістичною експертною інформацією (ЛЕІ) розуміють інформацію природною мовою (ПМ), отриману з текстів заданої предметної області (ПрО), шляхом лінгвістичного й експертного аналізу. У загальному випадку етап попередньої обробки можна охарактеризувати у вигляді технологічного ланцюжка у форматі: найменування етапу, підетапи, виконавець, вхідні/вихідні дані.

Експертна обробка полягає у виборі текстів, що належать однієї предметної області, класифікації текстів, складанні словників термінів та ін.

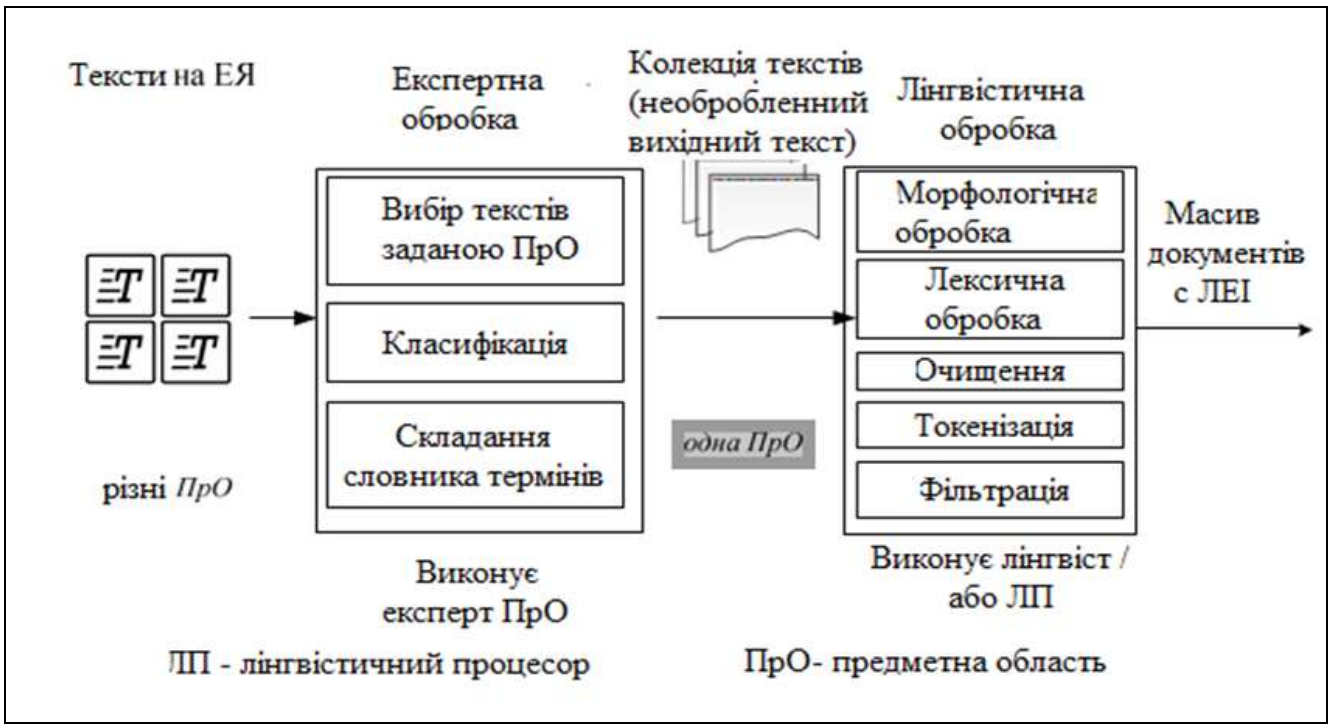

Рисунок 1.1 – Етапи обробки текстової інформації і одержання ЛЕІ [5]

Лінгвістична обробка підготовлених експертами колекцій текстів включає морфологічну, лексичну обробку, очищення тексту від «шумів» (наприклад, цифри, символи), токенізацію (розбивка його на окремо значимі одиниці, токени), фільтрація (наприклад, видалення стоп-слів) та ін. [5]. Результатом лінгвістичної обробки є масив документів, що полягають із термінів – значимих і нормалізованих слів предметної області. Отриманий результат є вхідними даними для модуля обробки й аналізу документів з ЛЕІ.

На другому етапі проводиться обробка й аналіз ЛЕІ, що отриманої з колекцій текстів заданої предметної області, з метою формалізації й інтеграції, яка вирішується за допомогою системи. Розглянемо модель обробки й аналізу документів з ЛЕІ, що дозволяє одержати формалізоване представлення ЛЕІ.

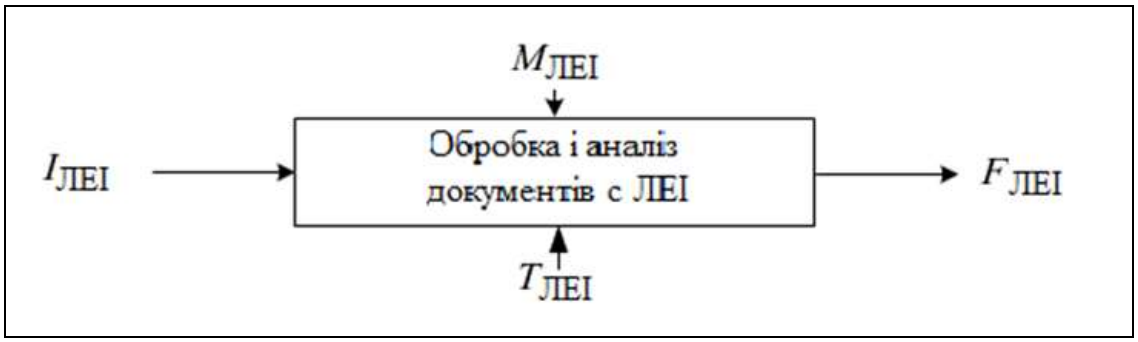

Рисунок 1.2 – Система обробки й аналізу ЛЕІ [5]

Обробка й аналіз вхідний ЛЕІ ($І_{ЛЕІ}$), отриманої з колекцій текстів, яка є неструктурованої, виконується за правилами формалізації, заданим деяким механізмом обробки інформації $F_{ЛЕІ}$. Під формалізацією розуміють процес представлення неструктурованої лінгвістичної експертної інформації у формалізованому виді. Множина керуючих впливів $М_{ЛЕІ}$ задане правилами обробки $Т_{ЛЕІ}$, які дозволяють одержати спеціалізований масив даних, необхідних для проведення подальших досліджень. Результатом аналізу й обробки ЛЕІ є формальне представлення $F_{ЛЕІ}$. У результаті формалізації ЛЕІ, отриманої з колекцій текстів знання, що витягають, здобувають явний вид і стають придатними для обробки інтелектуальними системами. Помітимо, що якщо формалізація текстів виконується на основі семантики, то результатом цього процесу є набір семантичних відносин, представлених, наприклад, у вигляді зв'язаного списку об'єктів – семантичного графа.

Вид формалізації (семантичні мережі, формальні логічні моделі, продукційні моделі) задається методами, моделями й алгоритмами інтелектуальної обробки ЛЕІ. Однією з перспективних форм представлення знань у формалізованому виді є онтологічний підхід [5]. Орієнтуючись на розробку інтелектуальних систем, цей підхід, дозволяє одержувати формалізовані представлення спеціальних предметних знань у зручному для комп'ютерної обробки виді, відокремлювати знання від програмного коду системи. Можливість для користувачів онтології додавати й змінювати об'єкти, атрибути й класи в міру уточнення цілей і завдань робить цей метод оптимальним у практичному використанні.

Онтологія – експліцитна специфікація концептуалізацій [6]. Під онтологією будемо розуміти модель представлення знань деякої предметної області у вигляді сукупності понять цієї предметної області й існуючих між ними відносин. Таким чином, онтології на базовому рівні повинні, насамперед, забезпечувати словник понять для представлення й обміну знаннями про предметну область, а також множина зв'язків, установлених між поняттями в цьому словнику.

У загальному випадку онтологія деякої предметної області формально може бути задана трійкою виду [7]:

$$Онт = < C, R, FA >,$$

де С – множина понять предметної області;

R – кінцева множина відносин між поняттями Про;

FA – кінцева множина функцій інтерпретації, заданих на поняттях і/або відносинах онтології О.

Множина відносин між поняттями предметної області R може включати ієрархічні (гіпоніми та гіпероніми, род-вид або частина-ціле) відносини, відносини еквівалентності (синоніми-антоніми), асоціативні відносини. За умови порожнечі множини відносин R і множини аксіом FA онтологія вироджується в глосарій. У випадку, коли множина відносин R полягає тільки з родовидових відносин, і множина аксіом F є порожнім, то онтологія вироджується в таксономію (ієрархічну структуру, що полягає з понять предметної області різних рівнів).

Використовуючи стосовно до аналізу й обробці текстів на ПМ підхід онтологічного моделювання, запропонований у роботі [7], процес формалізації ЛЕІ, отриманої з колекції текстів, можна представити послідовністю декількох незалежних етапів, на кожному з яких вирішується одне певне завдання, результати якої, у свою чергу, служать вихідними даними для завдання наступного, як правило, більш складного рівня. Згідно із цим підходом типова послідовність дій зводиться до наступних етапів, які аналогічні етапам автоматичної побудови онтології по колекції текстів.

Циклічність процесу: отримана онтологічна модель доповнюється новими поняттями й відносинами, оцінюється й потім уже використовується як база для подальшого розширення. Такий підхід використовується часто, і дозволяє витягти максимум інформації із вхідних колекцій текстів (див. рис. 1.3).

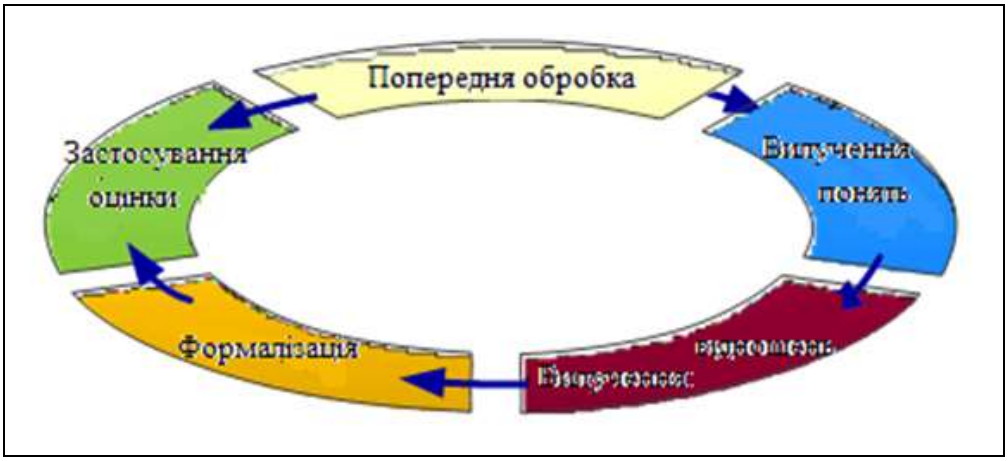

Рисунок 1.3 – Цикл витягу знань із колекції текстів і автоматичної побудови онтології [7]

Якість формованих онтологій багато в чому визначається повнотою обліку в онтологічній моделі найбільш значимих понять для колекції аналізованих текстів з обліком їх тематичної специфіки (під поняттями надалі розуміти найбільш значимі слова й словосполучення в аналізованому тексті, які можуть бути враховані в онтологічній моделі). Узагальнивши результати, виявлені наступні проблеми, що виникають на етапі витяги термінів і понять із текстів: а) не завжди вдається правильно знайти зв'язки між поняттями; б) не завжди вдається виділити поняття, що мають зв'язок з найбільшою кількістю інших понять; в) знайдені зв'язки між поняттями майбутньої онтології не завжди актуальні для конкретної предметної області.

Таким чином, нові методи обробки й аналізу ЛЕІ, отриманої з колекції текстів, повинні враховувати семантичні асоціації, засновані на смисловій моделі подібності слів, вирішувати завдання термінологічного покриття й побудови ієрархічних зв'язків між поняттями, а також ураховувати можливість інтеграції моделей знань. Елементи семантичного поля, тематичної й семантичною-семантичній-лексико-семантичної груп мають ієрархічну природу, виходить, родоподібні відносини є фундаментальними парадигматичними смисловими відносинами, за допомогою яких структурований словниковий склад мови» [8].

Тому основну увагу необхідно приділити методу витягу ієрархічної структури понять.

Відповідно до цього, для обробки й аналізу ЛЕІ, отриманої з колекцій текстів заданої предметної області, з метою формалізації й інтеграції, представляється можливим комплексний процес побудови простої онтології. Основною ідеєю є обробка ЛЕІ, витяг понять предметної області й побудова ієрархічної структури понять із можливістю багаторазового автоматичного поповнення отриманої моделі без вчителя (див. рис. 1.4).

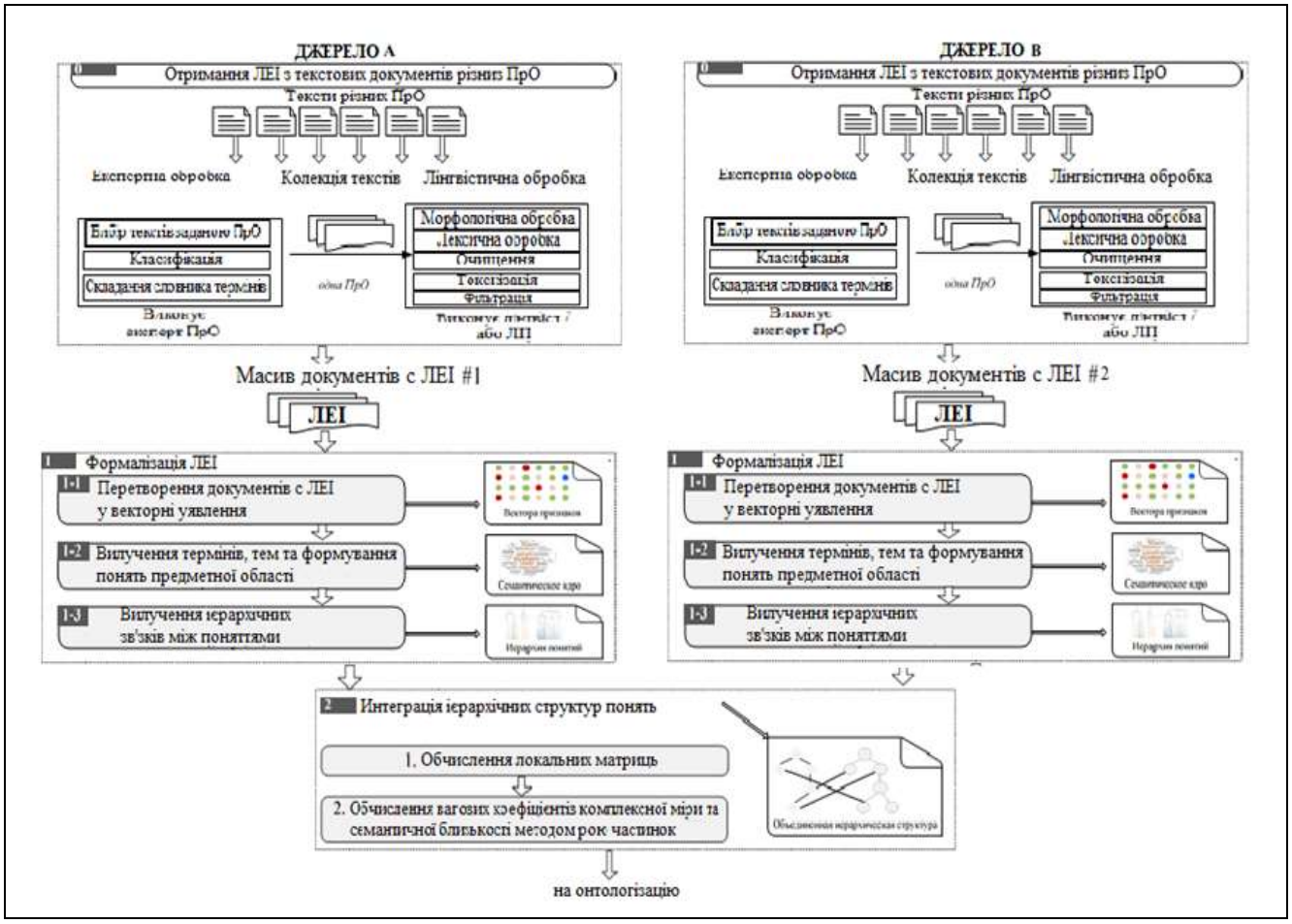

Рисунок 1.4 – Етапи формалізації і інтеграції ЛЕІ, отриманої з колекції текстів

Перелік етапів формалізації й інтеграції ЛЕІ:

– перетворення масиву ЛЕІ у векторні представлення. На цьому етапі масиви ЛЕІ, що мають вид послідовності слів (термінів), перетворяться до виду, з яким можуть працювати методи машинного навчання. Як правило, методи

машинного навчання працюють із векторами в багатомірному просторі ознак – ознаками є слова (терміни);

– витяг термінів і понять предметної області. Виділення термінів, термінів, понять із колекцій текстів є однієї зі складних підзадач формалізації ЛЕІ.

Термінами є одно- і багатокомпонентні лексичні групи, що відбивають зміст текстового документа. Термін – це ємні й важливі для окремо взятої предметної області слова в тексті, набір яких дозволяє одержати високорівневий опис його змісту для читача, забезпечивши компактне представлення й зберігання його змісту в пам'яті.

На відміну від термінів, які найчастіше багатозначні й несуть емоційне фарбування, терміни в межах сфери застосування однозначні й позбавлені експресії, вони не пов'язані з контекстом.

Термінами є слова або словосполучення, що є назвою поняття деякої предметної області. Вони виступають, що спеціалізують, точними позначеннями, характерними для конкретної сфери предметів, явищ, їх властивостей і взаємодій.

Для виділення з текстів термінів понять найчастіше використовуються семантичні й статистичні критерії. При використанні семантичних критеріїв пошуку термінів і понять виконується логіко-семантичне співвіднесення професійного поняття й відповідного йому язикового вираження, а потім виконується перевірка його на термінологічність шляхом застосування логічних теорем.

Застосування статистичних методів дозволяє об'єктивно й з більшою точністю відібрати терміни в спеціальних текстах.

Семантичний критерій виконується тільки людиною.

Витяг ієрархічних зв'язків між поняттями. Терміни, що формують корпус досліджуваної предметної області, структуруються за допомогою попередньо сформованої системи понять предметної області [6]. Основним відношенням, що враховуються при побудові онтології, є родовидове відношення між поняттями (відношення гипоним та гипероним), на основі якого формується таксономія понять.

Гіпоним – поняття, що виражає приватну сутність стосовно іншого, більш загального поняття.

Гіперонім – слово з більш широким значенням, що виражає загальне, родове поняття, назва класу (множина) предметів. Гіперонім є результатом логічної операції узагальнення. Наприклад, терміни «software development» і «software» зв'язані між собою, оскільки термін «software» є гіперонимом стосовно «development» і виражає загальне поняття -«програмне забезпечення», а термін «software development» (розробка ПО) має більш «вузьке» значення, і позначає елемент цієї множини, тобто виступає гіпонімом [10].

Інтеграція ієрархічних структур понять дозволяє розширити отриманий опис предметної області шляхом об'єднання формалізованих представлень різних колекцій текстів. Інтеграцію онтологічних моделей визначено як процес знаходження подібності двох онтологій O_1 і O_2 і, як результат, створення нової онтологічної моделі 03, що поєднує, що й погодить семантичні представлення вихідних онтологій. При інтеграції ієрархічних структур виникає ряд проблем, пов'язаних з різного роду невідповідностями. Основними типами невідповідності схем даних можуть бути конфлікти іменування (у різних схемах використовується різна термінологія), семантичні конфлікти (обрані різні рівні абстракції для моделювання), різна глибина таксономічних зв'язків.

Таким чином, основними завданнями, пов'язаними з інтелектуальним аналізом текстової й лінгвістичної експертної інформації, є:

– представлення ЛЕІ у векторній формі;

– витяг термінів і понять;

– витяг ієрархічних зв'язків і побудова ієрархії понять;

– поповнення бази знань за допомогою інтеграції ієрархічних структур колекцій текстів, отриманих з різних джерел.

Перераховані етапи аналізу й обробки ЛЕІ розглядаються в якості початкового етапу автоматичної побудови онтологічної бази знань. Проведемо огляд відомих методів аналізу й обробки ЛЕІ, що дозволяють формалізувати

неструктуровану ЛЕІ й забезпечити інтеграцію формалізованих структур різних джерел.

1.2 Аналіз методів обробки ЛЕІ

Огляд існуючих методів інтелектуального аналізу текстів і обробки ЛЕІ проведемо з погляду позначених у попередньому параграфі завдань: перетворення документів з ЛЕІ у векторні представлення, витяг термінів і понять предметної області, витяг ієрархічних зв'язків між поняттями, інтеграція ієрархічних структур понять, а також проведемо аналітичний огляд наукових праць, присвячених проблемам формалізації й інтеграції знань, що втримуються в колекції текстів.

Основою рішення більшості завдань, пов'язаних з інтелектуальною обробкою неструктурованої текстової інформації, у т.ч. ЛЕІ, є векторна модель – представлення окремо взятого документа або колекції документів як вектора у векторному просторі.

Часто використовуваної при рішенні завдання векторної представлення текстів є гіпотеза «мішок слів», згідно з якою текст являє собою неупорядкований набір слів. Однієї з різновидів моделі «мішок слів» є модель «мішок термінів». У моделі «мішок термінів» основним об'єктом є терм, який має єдиний атрибут – частоту зустрічальності в тексті. Термінами є слова або словосполучення, що несуть інформацію про тематику документа. Термін, або ключові слова являють собою короткий опис документа.

У якості представлення документів використовується підхід «мішок термінів». У такому підході значущими вважаються не всі слова й словосполучення, а тільки певний набір заданих заздалегідь термінів. При цьому, побудувавши матрицю входжень термінів у документи, з ними можна працювати точно так само, як зі звичайними словами в описаній моделі «мішок слів». Така модель із використанням термінів корисна у випадку, коли документи є не

більшими й належать вузькій предметній області. Наприклад, запропоновано використовувати таку модель для кластеризації анотацій наукових статей.

Колекція документів, представлена моделлю «мішок слів», відображається в числову матрицю A, рядки якої відповідають термам, а стовпці – документам (документна матриця). Елементами матриці A є вагові коефіцієнти (ваги), які можуть бути обчислено одним з наступних способів:

– бінарний – ознака приймає значення 1, якщо в документі зустрічається відповідна *n*-грама, і 0 – якщо інакше;

– по кількості входжень – таке представлення містить тільки термін, які належать документу й частоту появи цих слів у документі;

– на основі зворотної частоти (TF-IDF, Term Frequency – Inverse Document Frequency) – представлення документів у даному просторі полягає у формуванні вектора частот слів (термінів), помноженого на вектор зворотної частоти слів у кожному документі. Параметр IDF (зворотна частота зустрічальності в корпусі) указує на загальну зустрічальність слова у всім корпусі. Отже, поява в документі слів, що часто зустрічаються, нормалізується. Високі ваги будуть у термінів, які зустрічаються в документі часто, але рідко у всій колекції;

– метод BM25 – обмежується значимість частоти n-грами, а також вона не тільки нормалізується по його розміру, але й обмежується зверху, що дозволяє уникнути присвоювання слову занадто великої ваги.

Залежно від розв'язуваного завдання можуть застосовуватися різні способи обчислення вагових коефіцієнтів термінів, які прийнято називати ваговими схемами.

Основним недоліком частотних методів є висока розмірність і розрідженість одержуваних векторів (обумовлене обсягом використовуваного словника). Із цієї причини представлення TF – IDF часто комбінується з методами зниження розмірності. Тому вживають спроби змінити простір TF – IDF на простори меншої розмірності, вектори яких містять уже частково оброблену статистичну інформацію.

Методи представлення текстової інформації, засновані на матричному розкладанні, дозволяють знизити розмірність простору. Одним з методів даної групи є латентно-семантичний аналіз (Latent Semantic Analysis, LSA), який являє собою сингулярне розкладанні (singular value decomposition, SVD) матриці терм-документ. За допомогою Svde-розкладання матриця розкладається на трійку матриць, лінійна комбінація яких є досить точним наближенням до вихідної матриці. Переваги методу – оптимальність матричного розкладання матриці частот слів у документах. Недоліками цього методу є відносно висока обчислювальна вартість одержання підпростору, а також сильне допущення про нормальність помилки при розкладанні матриці документ-терм.

Методи, засновані на ідеях дистрибутивної семантики, це методи знаходження низкорозмірного векторного простору слів, засновані на методах ковзаючих по тексту контекстних вікон при складанні матриці терм-терм. Найбільш популярні методи даного класу – word2vec, Glove і fasttext. Перші два методи розкладають матрицю терм-терм на дві матриці основну й контекстну. Word2Vec використовує для цього мінімізацію логарифма правдоподібності, узятого з негативним знаком, а Glove – мінімізацію зваженої суми квадратів помилок. В відмінність від перших двох методів Fasttext використовує додаткову інформацію про морфологію слова при побудові вектора слова, представляючи його (слово) виді суми буквених n-грам. Перевагою методу Fasttext є те, що він є більш швидким, чому традиційні методи глибокого навчання. Недолік – відсутність інтерпретованості координат отриманих векторів. Інтерпретованість даного варіанта можлива тільки щодо міри подібності між векторами.

Побудова ієрархії можливо шляхом узагальнення й виявлення подібності елементів по яких-небудь ознаках і властивостям. У цей час розроблене велику кількість методів ієрархічної кластеризації. У рамках справжньої роботи проведений огляд тих методів, які підходять для рішення завдання витягу ієрархічних зв'язків між поняттями. Аналіз літератури показав, що існують різні методи витягу ієрархії понять (термінів) предметної області.

Ієрархічні методи кластеризації добре зарекомендували себе серед статистичних методів обробки текстової інформації. При ієрархічної виконується послідовне об'єднання менших кластерів у більші (агломеративні) або поділ більших кластерів на менші (дивизимні) на підставі ступеня близькості між ними.

У таблиці 1.1 представлений короткий опис розглянутих методів перетворення текстової інформації у векторну форму, заснованих на гіпотезах «мішок слів» і «мішок термів».

Таблиця 1.1 – Порівняльний огляд методів представлення текстової інформації у векторну форму

Метод векторизації	Принцип дії	Найбільш підходяще застосування	Недоліки
Частотний	Підрахунок частоти входження слів/термів	Баєсівські моделі	Саме слова, що часто не зустрічаються, завжди є самими інформативними
Бінарна представлення	Визначення логічної ознаки присутності слова/терма	Нейронні мережі	Усі слова рівновіддалені, тому нормалізація є важливою
TF-IDF	Нормалізація частоти слів по документах	Додатки загального призначення, у т.ч. BTM	Помірковано часто слова, що зустрічаються, можуть бути не репрезентативними для теми документа
Дистрибутивна семантика	Кодування подібності лексем на основі контексту	Моделювання складних відносин	Великий обсяг обчислень, складність масштабування без застосування додаткових інструментів

1.3 Методи витягу термів, тем і понять

У цей час питання розробки методів автоматичного виділення слів і термів з текстів досить активно розглядаються як у вітчизняних, так і в закордонних роботах. У роботі [30], показане, що розроблені методи даної групи засновані, переважно, на побудову частотних словників, конкордансів (словників словосполучень) та ін. Враховуючи результати, представлені в [14], були виділені наступні класи методів виділення термів, які представлено на рис. 1.5.

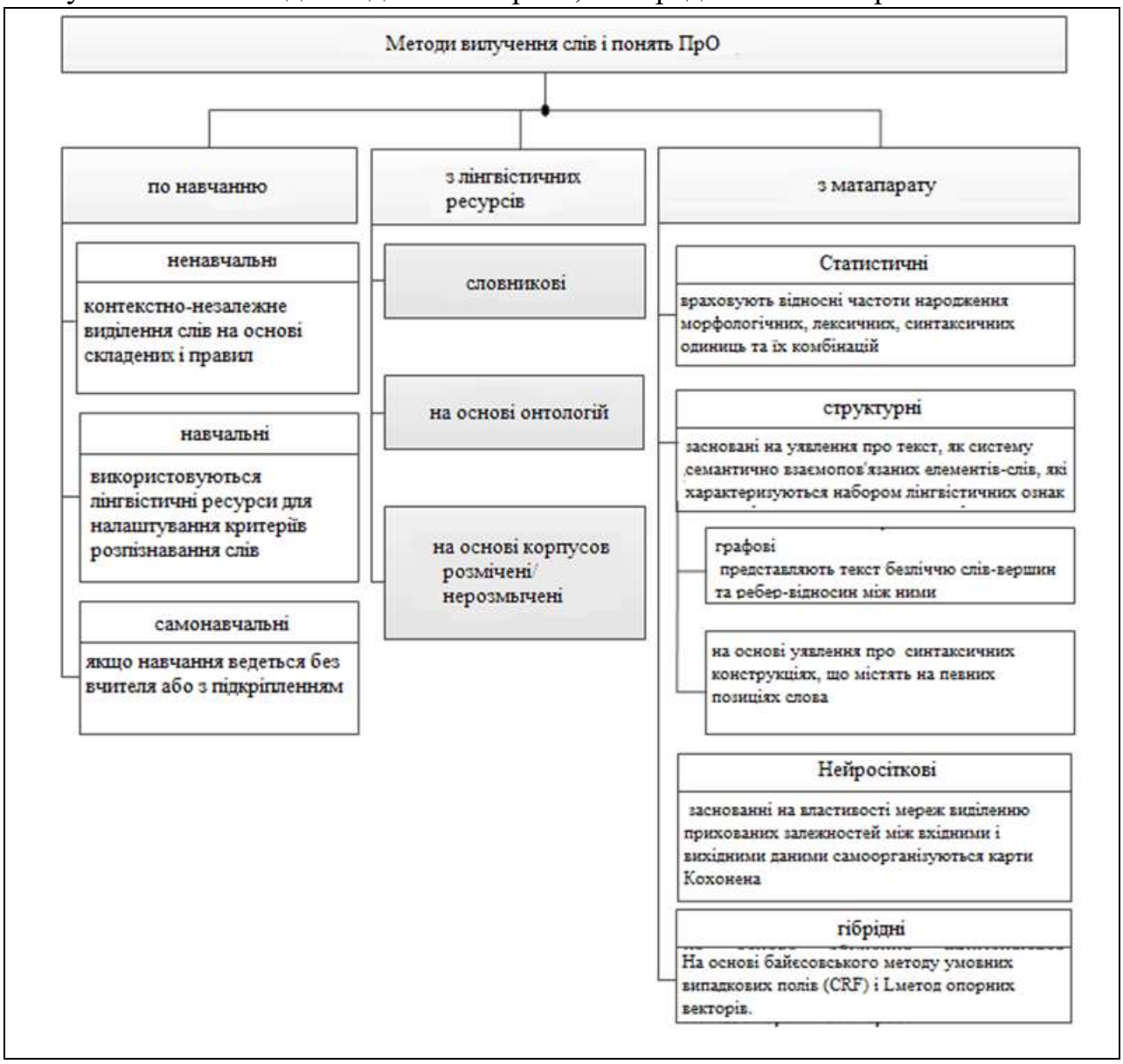

Рисунок 1.5 – Методи виділення термів [14]

Незважаючи на триваюче вдосконалювання класичних статистичних і структурних алгоритмів витягу термів, акцент розроблювачів змістився в область гібридних рішень із навчанням на основі текстових корпусів.

Група статистичних методів, що працюють із моделями ймовірносно-тематичного моделювання (ВТМ), оскільки вони дозволяють витягати семантично близькі термін, контекстно-залежні значення лексичних одиниць, визначати можливі тематики документів. Враховуючи результати, отримані в [14], [15], був зроблений висновок, що ці методи можна використовувати для виділення не тільки термів, але й понять предметної області. В основі методів ВТМ лежить стохастичне матричне розкладання, матриці терм-документ на матриці терм-тема й тема-документ, а також теорія імовірності, зокрема наївний баєсівський класифікатор [16]. Найпоширенішими методами групи ВТМ є імовірнісний латентно-семантичний аналіз (ВЛСА) [17] і модель латентного розміщення Дирихлє (ЛРД) [18].

Метод ВЛСА заснований на виявленні зв'язків між схованими (латентними) змінними – темами з кожною спостережуваною змінною -словом або темою. Завдання полягає в знаходженні схованих змінних. Документ, згідно ВЛСА, може відноситися до декількох тем із деякою ймовірністю, що є відмінною рисою цієї моделі в порівнянні з походами, не заснованими на імовірнісному моделюванні. Перевагами методу ВЛСА є гарна якість визначення тематик у великих корпусах текстів, можливість знаходження схованих семантичних залежностей між словами.

Експериментальні дослідження ВЛСА показали, що цей метод схильний до перенавчання і не дозволяє природно додавати в модель нові документи [19]. Ефект перенавчання призводить до того, що побудована модель добре пояснює приклади з навчальної вибірки, але відносно погано працює на прикладах, що не брали участь у навчанні, тобто в навчальній вибірці виявляються деякі випадкові закономірності, які відсутні в генеральній сукупності. Іншими недоліками ВЛСА є висока обчислювальна складність і низька швидкість роботи, що вимагає повторного обчислення всіх метрик для всього корпуса у випадку додавання нового документа,

високі вимоги до корпуса текстів, який повинен складатися з множини різноманітних по тематиках текстів.

Розвитком методу ВЛСА є метод латентного розміщення Дирихле (ЛРД), який доповнений припущенням про те, що розподілу тем по документах і слів по темах породжуються розподілами Дирихле. Метод ЛРД дозволяє піти від недоліків імовірнісного ЛСА, таких як «перенавченість» і відсутність закономірності при генерації документів з набору отриманих тем за рахунок застосування байєсівської регулярізації – додавання деяких апріорних розподілів на параметри моделі [13].

Відома множина модифікацій моделі ЛРД, що враховують специфіку текстових колекцій. Короткий огляд інших моделей, створених на основі моделей ЛСА й ЛРД, що враховують іншу інформацію, що дозволяє поліпшити результати формування тем тексту, наведено в таблиці 1.2.

Таблиця 1.2 – Огляд основних моделей тематичного моделювання

№	Модель	Опис
1.	Спільна імовірнісна модель (joint probabilistic model)	Модельо-узагальнення імовірнісної ЛСА, що враховує не тільки тематики, але й природній зв'язок між документами
2.	Автор-Тематична модель (author-topic model)	враховуючі інформацію про автора тексту
3.	Модель ЛРД із використанням словника Word-Net (latent Dirichlet allocation with Wordnet, LDAWN)	на основі ЛРД, що враховують контекст слова; замість породження слів прямо з теми, кожній темі був протипоставлений «шлях» через ієрархічний словник Word-Net
4.	Динамічна тематична модель (dynamic topic model, DTM)	ураховуються зміни, що відбуваються з тематикою текстів у корпусі із часом
5.	Схована тематична марковська модель ЛРД (Hidden Markov Model with Latent Dirichlet Allocation, HMM-LDA)	які дозволяють розглядати текст як деяку сукупність біграм, n-грам і навіть пропозицій або абзаців.
6.	Двомовна тематична модель ЛДА (bilingual model with latent Dirichlet allocation, Bilda)	заснована на дистрибутивній гіпотезі, що говорить, що лінгвістичні одиниці, що зустрічаються в схожих контекстах, мають близькі значення
7.	Адитивна регуляризація (additive regularization for topic modeling)	заснована стандартної імовірнісної ЛСА моделі й уведенні додаткових критеріїв-регуляризаторів

Інші відомі методи відрізняються в основному способом згладжування частотних оцінок ймовірностей, причому всі вони стають практично еквівалентними, якщо оптимізувати гіперпараметри апріорних розподілів, як пропонується в [133].

У результаті розгляду деяких основних методів виділення термів і понять предметної області можна дійти до висновку, що методи, засновані на ймовірносно-тематичному моделюванні (зокрема, модель латентного розміщення Дирихле) щонайкраще придатні для рішення поставленого завдання – виділення термів і понять предметної області [20]. Основним недоліком моделі ЛРД (оскільки спочатку вона розроблялася для рішення завдань зняття неоднозначності, інформаційного пошуку, порівняння текстів) стосовно до завдання виділення понять є те, що темам, що витягають, не привласнюються імена (тільки індекси), що суттєво впливає на ефективність рішення завдання.

1.4 Методи витягу ієрархічних зв'язків між поняттями

Побудова ієрархії можлива шляхом узагальнення й виявлення подібності елементів по яких-небудь ознаках і властивостям [21]. У цей час розроблено велика кількість методів ієрархічної кластеризації.

Проведений огляд тих методів, які підходять для рішення завдання витягу ієрархічних зв'язків між поняттями. Існують різні методи витягу ієрархії понять (термів) предметної області. До найпоширеніших можна віднести наступні методи, які показано на рисунку 1.6.

Перевагами розглянутих методів є: гарна інтерперетованість результатів, прийнятна якість знаходження відносин.

Оскільки в справжній роботі розглядається можливість побудови ієрархії термів без вчителя, найбільш підходящими є методи статистичної групи [21].

Рисунок 1.6 – Методи витягу ієрархічних зв'язків між поняттями [21]

Статистичну групу методів, що дозволяють витягати ієрархічні зв'язки між поняттями, утворюють ієрархічні імовірнісні методи й методи ієрархічної кластеризації. Ієрархічні імовірнісні моделі дозволяють ураховувати й виявляти ієрархічну структуру тем у колекції документів. Для цього робляться апріорні припущення про тип розподілів підтем і слів у темах більш високого рівня. Ці припущення визначають властивості ієрархічної моделі.

Недоліками ієрархічних імовірнісних моделей є: робота за принципом «чорний ящик», складність реалізації, висока чутливість до помилок на верхніх рівнях (наприклад, дублювання або змішання тем), не використовуються для побудови ієрархічних відносин понять онтології.

Ієрархічні методи кластеризації добре зарекомендували себе серед статистичних методів обробки текстової інформації [22]. При ієрархічній кластеризації виконується послідовне об'єднання менших кластерів у більші (агломеративні) або поділ більших кластерів на менші (дивизимні) на підставі ступеня близькості між ними.

Таблиця 1.3 – Огляд ієрархічних імовірнісних моделей

Назва моделі	Структура у вигляді спрямованого ациклічного графа	Внутрішні вузли генерують слова	Добір числа кластерів на рівні	Добір числа рівнів
Ієрархічна модель латентного розміщення Дирихлє (hlda)	не реалізоване	реалізоване	реалізоване	не реалізоване
Ієрархічна модель Пачинко HPAM	реалізоване	реалізоване	не реалізоване	не реалізоване
Непараметрична модель Пачинко NPBPAM	реалізоване	не реалізоване	реалізоване	не реалізоване

Початкові кроки в процедурах ієрархічного агломеративного кластерного аналізу ідентичні. Для визначення відстані між двома об'єктами, що належать різним кластерам, можуть застосовуватися правила:

– одиночному зв'язку: відстань між двома кластерами Clust, Clust′ визначається по відстані між двома найбільш близькими об'єктами (найближчими сусідами);

– повному зв'язку: відстані між кластерами Clust, Clust′ визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах;

– незваженого середнього – відстань між кластерами Clust, Clust′ обчислюється як середня відстань між усіма парами об'єктів у них.

По обчислених відстанях виконується об'єднання кластерів: стовпці й рядка в матриці відстаней між об'єктами для близьких об'єктів віддаляються, а на їхнє місце вставляються нові стовпець і рядок з перерахованими значеннями об'єктів, щоб не порушити діагональ нулів.

В результаті розмір матриці наближення зменшується на одиницю, а найменше значення стає параметром дендрограми, тому що визначає відстань між об'єктами й указує номер групи n + 1. Кожний наступний крок має злиття між двома об'єктами або між об'єктом і групою або між двома групами, для яких ступінь близькості мінімальний, потім виконується аналогічне перерахування, і

об'єднання в групи. Процедура завершується, коли розмірність матриці близькості дорівнює 2 × 2. Загальним недоліком традиційних алгоритмів є нездатність виділяти кластери довільної форми. Ці алгоритми базуються на припущенні про сферичну форму кластерів, що найчастіше приводить до помилок кластеризації або зайвої роздробленості інформаційних класів.

Агломеративні ієрархічні алгоритми дозволяють працювати з більшим обсягом інформації, компактно й наочно представляють розглянуті дані, при стиску даних скорочується обсяг збережених даних. Показане, що методи ієрархічної кластеризації можна застосовувати для завдання побудови ієрархічних відносин між поняттями предметної області Для автоматичного виявлення ієрархічних зв'язків між поняттями застосовується агломеративна кластеризація й використовується великий корпус текстів для витягу статистичної інформації про спільну зустрічальність слів.

1.5 Обґрунтування цілей дослідження

Застосування методів онтологічного моделювання при розробці методів формалізації ЛЕІ в інтелектуальних системах обробки текстової інформації представляється перспективним підходом. Основними труднощами, пов'язаними з представленою ЛЕІ є: виділення понять предметної області й побудова ієрархії понять предметної області. Методи машинного навчання, особливо – методи навчання без вчителя, дозволяють виявляти сховані закономірності в неструктурованій інформації в автоматичному режимі. Застосування таких методів дозволяє підвищити повноту доступних семантичних ресурсів. Існуючі методи автоматичного й автоматизованого побудови семантичних мереж припускають або наявність доступних готових ресурсів високої якості для

інтеграції, або формують тільки поняття, або формують тільки семантичні відносини.

У результаті аналізу робіт був зроблений висновок про необхідність розробки методів аналізу й обробки ЛЕІ, що дозволяє усунути виявлені недоліки розроблених методів і сполучити використовувані методи в одному процесі.

Формалізація й інтеграція знань, що втримуються в колекціях текстів, являє собою нетривіальне завдання, має всі особливості слабоструктурованих проблем. Розглянуті методи характеризуються складністю опису моделей для генерації коду й високими кваліфікаційними вимогами до користувача; відсутністю можливості спільної роботи користувачів; часто залежать від предметної області, мови, додатка й т.п., що ускладнює або унеможливлює їхній перенос на інші предметні області, мови, додатки. Наукові дослідження в даній області характеризуються тим, що звичайні способи обробки інформації стосовно до обробки більших обсягів неструктурованих даних не забезпечують необхідної швидкості, повноти і якості її переробки.

Таким чином, показана необхідність розробки більш точних методів аналізу й обробки ПМ, що враховують нелінійну й ієрархічну природу тексту, що дозволяють одержувати структуровану представлення текстових документів. У цьому зв'язку запропоновано вдосконалити базові методи аналізу й обробки ЛЕІ, у т.ч. метод тематичного моделювання, метод ієрархічного кластерного аналізу й метод семантичної інтеграції даних.

Необхідна розробка аналітичного методу формалізації лінгвістичної експертної інформації, отриманої з колекції текстів заданої предметної області, у вигляді ієрархічної структури понять, яка складається з підзадач:

– векторна представлення масиву ЛЕІ;

– витяг термів і тем, формування понять предметної області;

– побудова ієрархічних зв'язків між поняттями.

Аналіз методів векторної представлення ЛЕІ показав, що:

– модель «мішка слів» добре підходить для завдань тематичного моделювання; наявний недолік – висока розмірність – буде скомпенсований моделлю ЛРД, яка дозволяє зменшити розмірність простору;

– модель латентного розміщення Дирихле добре підходить для тематичного аналізу тексту, виділення семантично близьких термів і тем, і відображення отриманих тем у простір понять;

– для виділення ієрархічних відносин – побудови ієрархічної структури понять предметної області добре підходить метод ієрархічної агломеративної кластеризації, оскільки дозволяє працювати з більшим обсягом інформації, компактно й наочно представляє кластеризовану інформацію, дозволяє будувати ієрархію множина об'єктів.

Базова ймовірносно-тематична модель, заснована на латентному розміщенні Дирихлє дозволяє виявити в тексті слова й теми, які не мають назв (тільки індекси). У результаті інтерпретація тем і формування понять предметної області виконується експертами вручну. Тому необхідно модифікувати метод ймовірносно-тематичного моделювання таким чином, щоб теми, що витягаються, іменувалися в автоматичному режимі – включити в модель підрівень автоматичного формування понять. Це дозволить зі знайдених термів і тем формувати поняття, що описують предметну область. Для знаходження ієрархічних зв'язків і побудови ієрархічної структури понять до отриманих понять застосувати ієрархічний кластерний аналіз.

Аналіз методів інтеграції даних показав, що інтеграція ієрархічних структур, побудованих по різних колекціях текстів однієї предметної області, може бути вирішена шляхом обчислення комплексного міри семантичної близькості.

Необхідно розробити алгоритм інтелектуальної обробки лінгвістичної експертної інформації для формалізації й інтеграції колекцій текстів заданої предметної області.

Алгоритм повинен бути розроблений на основі запропонованих методів і дозволяти створювати процедуру автоматичної побудови ієрархічних структур декількох колекцій текстів.

Реалізувати запропоновані методи й алгоритм аналізу й обробки лінгвістичної експертної інформації, провести оцінку їх застосовності для автоматичної побудови простої онтології. В результаті рішення розглянутих завдань повинні бути отримані методи й алгоритми обробки ЛЕІ і її формалізації, що дозволяють виділяти слова, словосполучення й теми з документів, що характеризують задану область знання, створювати поняття предметної області, будувати ієрархічну структуру понять предметної області (просту онтологію), здійснювати інтеграцію даних, отриманих з різних джерел.

2 МЕТОДИ АНАЛІЗУ ЕКСПЕРТНОЇ ІНФОРМАЦІЇ

2.1 Методи інтеграції формалізованих структур

Існуючі технології інтеграції даних забезпечують злиття й відображення даних на фізичному, логічному й семантичному рівнях. На сьогоднішній день у неоднорідних ІС накопичується значний обсяг знань, і при інтеграції таких систем виникає проблема систематизації й структурної представлення знань про різні предметні області. При рішенні проблеми інтеграції даних використовувалися онтології. Онтологічний підхід забезпечує новий рівень у рішенні завдань інтеграції інформації, отриманої з різних джерел.

Для забезпечення семантично коректного взаємозв'язку неоднорідних інформаційних систем необхідно зіставити онтології, що лежать у їхній основі, і з'ясувати їх спільності і відмінності. Це завдання вирішується шляхом оцінки семантичної близькості понять онтологій.

Для забезпечення семантично коректного взаємозв'язку неоднорідних інформаційних систем необхідно зіставити онтології, що лежать у їхній основі, і з'ясувати їх спільності і відмінності. Це завдання вирішується шляхом оцінки семантичної близькості понять онтологій [21], [22].

Труднощі порівняння різних онтологій предметних областей полягають в відмінності імен понять і відносин, а також підходах до визначення понять.

При відображенні двох онтологій виконується пошук для кожного поняття однієї онтології подібного його поняття іншої онтології.

Основним недоліком більшості методів визначення семантичної близькості є необхідність залучення експерта для підтвердження коректності виявлення подібностей і відмінностей семантичних понять [22].

Усунути суб'єктивність і трудомісткість процесу обчислення семантичних заходів близькості між поняттями дозволяють генетичні й еволюційні методи. Огляд найбільш перспективних підходів, з погляду розв'язуваного в роботі завдання представлено в таблиці 2.1.

Таблиця 2.1 – Огляд методів автоматичного знаходження міри близькості між поняттями онтології

№ пп	Підхід
1.	Розглянута реалізація рішення завдань природно-мовної обробки наукового тексту, здійснена із застосуванням технологій генетичного й автоматного програмування, яка дозволила створити технологію рішення завдань побудови онтологій, що відрізняється майже повною автоматичною обробкою.
2.	Генетичний алгоритм, який заснований на використанні аналогів з еволюційними процесами репродукції, кросінговера, мутації й природного добору.
3.	Модель вирівнювання онтології була представлена як завдання знаходження оптимального мінімального значення, у якій цільова функція заснована на нечіткій близькості.
4.	Модифікований генетичний алгоритм, що дозволяє визначити різні метрики близькості в одну. Імітаційний підхід, що представляє комбінацію еволюційного й локального пошуку методів, дозволяє одержати кращі результати, у порівнянні з генетичним алгоритмом.
5.	Гібридний підхід, заснований на генетичному алгоритмі, який визначає оптимальну конфігурацію для відображення онтологій (використовує знання Wordnet).
6.	Еволюційний підхід до рішення завдання інтеграції множинних онтологій для забезпечення сумісності й репрезентації даних і знань в інтелектуальних інформаційних системах.
7.	Дискретний метод оптимізації роя часток . Недоліком запропонованого методу є те, що кожний елемент у положенні частки повинен бути призначений своїй власній частковій відповідності, що ускладнює застосування даного підходу.

У роботах [23], [24], запропоноване рішення завдання обчислення комплексного міри семантичної близькості між поняттями із застосуванням методу «роя часток». Метод «роя часток» (МРЧ) є евристичним методом оптимізації, що використовують імітацію соціальної поведінки роя часток [21]. Агентами в МРЧ є частки, які в кожний момент часу мають у просторі параметрів завдання деяке положення й швидкість. Правила, по яких частка міняє своє положення й швидкість, визначаються на основі обчислення цільової функції частки. Експериментальні дослідження підтвердили, що метод роя часток

дозволяє знаходити квазіоптимальні рішення за прийнятний час. Запропоновані підходи дозволили підвищити точність і оптимальність знаходження рішення.

2.2 Аналіз аналітичних методів формалізації

Описано аналітичні методи формалізації ЛЕІ, отриманої з колекції текстів заданої предметної області, у вигляді ієрархічної структури понять, і інтеграції отриманих ієрархічних структур понять (ІСП) декількох колекцій текстів, що належать до заданої предметної області. Принциповою відмінністю пропонованих методів є системний підхід, основу якого становить сукупність методів тематичного й онтологічного моделювання, машинного навчання й еволюційних методів. Це дозволяє одержувати якісно нові наукові результати, а також скоротити час пошуку прийнятних рішень для поставлених завдань, усунути суб'єктивність описів понять онтології залежно від різних точок зору розроблювачів онтологій.

Формалізація ЛЕІ полягає в побудові моделі представлення знань у вигляді простої онтології – ІСП, або дендрограми. Вузлами дендрограми є поняття. Тому завдання зводиться до витягу з колекцій текстів понять і ієрархічних зв'язків між ними. Для розробки методу використовуються методи ймовірносно-тематичного моделювання й ієрархічної агломеративної кластеризації, основною областю застосування яких є класифікація й кластеризація текстових документів. Обґрунтуванням застосування ймовірносно-тематичного моделювання і ІАК до завдання витягу понять і формування ієрархічних зв'язків між ними є наступне:

– семантичні поля понять (тема) характеризуються сукупністю язикових одиниць;

– усі слова (термін), що входять у семантичне поле, конкретизують одне загальне поняття;

– семантичні поля організовані по ієрархічному принципу – вони відбивають родо-видові відносини між поняттями, що відображають предмети і явища дійсності.

Аналітичний метод формалізації ЛЕІ полягає в послідовному виконанні наступних етапів.

На першому етапі аналітичного методу формалізації ЛЕІ вирішується підзадача 1 – одержання векторної представлення документів з ЛЕІ. На вхід аналітичного методу формалізації ЛЕІ з попередньої обробки надходить масив документів з ЛЕІ (далі по тексту «документи») – тексти, розбиті на термін. Прийнявши гіпотезу «мішок термінів», виконується обчислення числа входжень термінів у колекцію документів і формування матриці терм-документ.

На другому етапі аналітичного методу формалізації ЛЕІ вирішується підзадача 2 – витяг термінів, тем і понять предметної області. Вихідними даними підзадачі 2 є терм-документна матриця, отримана в результаті рішення підзадачі 1. По терм-документній матриці на основі моделі ЛРД виконується обчислення матриці імовірнісних розподілів термінів по темах і матриці розподілу тем по документах. Отримана матриця тем із застосуванням словника термінів предметної області, заданого експертом, відображається в простір (вектор) понять.

На третьому етапі аналітичного методу формалізації ЛЕІ вирішується підзадача 3 – витяг ієрархічних зв'язків між поняттями. Вихідними даними подзадачі 3 є вектор понять, який є результатом подзадачі 2. Між поняттями обчислюються відстані з обліком їх імовірностей, і формується матриця попарних зв'язків понять (матриця відстаней). По отриманій матриці виконується ієрархічний кластерний аналіз, будується дендрограма – бінарне дерево, яке описує структуру вкладеності з ребрами (імовірності) між кластерами понять.

Отримана терм-документна матриця D використовується для рішення наступної підзадачі – витяг термінів, тем і понять ПрО.

Витяг термінів, тем і понять предметної області (підзадача 2). Рішення підзадач витягу термінів, тем і понять засноване на побудові тематичної моделі множини документів.

Знайдемо матриці тем у документах і матриці термінів у темах. Основною ідеєю ВТМ на основі ЛРД є те, що документи представляються сумішшю розподілів схованих тем t з множини можливих □ (□ ∈ □), де кожна тема визначається імовірнісним розподілом на множину термінів [8].

Тематичним моделюванням називається рішення зворотного завдання в наступному формулюванні.

Оцінка шуканих параметрів $\boldsymbol{\Phi}$ та $\boldsymbol{\Theta}$ може бути виконана за допомогою формули згорнутого семплювання Гиббса.

Взначається вектор понять ПрО. Для знаходження вектора понять ПрО необхідно векторний простір знайдених тем у документах відобразити в простір термінів. Оскільки термін – це слово або словосполучення, яке є назвою деякого поняття, після відображення тем у простір термінів одержимо поняття.

Застосуємо зовнішнє джерело знань – словник термінів, відповідний до термінології предметної області. Словник термінів ПрО являє собою множину $\mathbb{C}$, яка складається з векторів термінів $C_p \in \mathbb{C}$.

За фактом установлення еквівалентності термінів і понять для кожної теми ставиться в мітка, відповідна до назви еквівалентного їй терміна ($C_{екв}$, абсолютна відповідність).

У випадку рівності заходів близькості однієї теми (для якої вже встановлена абсолютна відповідність) із двома різними термінами Ср й C_{p+1}, для наступної пари Ср+1 і t_k повторно виконується аналіз значень заходів близькості між цим терміном C_{p+1} і іншими темами за винятком теми, для якої абсолютна відповідність була встановлена. Вибирається максимальне значення міри близькості й виконується перевірка еквівалентності відповідно до граничного значення δ. За фактом установлення еквівалентності терміна й поняття для теми

ставиться в мітка, відповідна до назви еквівалентного їй терміна (Секв, часткова відповідність).

У випадку відсутності еквівалентності межу терміном і темою, створюється нове поняття. Із цією метою проводиться оцінка ймовірностей розподілу термінів W_j у темі t_k. Вибирається терм із максимальним значенням імовірності p(w|t), і для цієї теми ставиться у відповідність мітка, відповідна до назви терма ($w_{екв}$, умовна відповідність).

У відповідність зі сказаним виконується відображення вектора термінів Ср і матриці розподілу термінів Ф у вектор понять Y. Довжина вектора dim(Y) = NT. Елементами вектора є строкова змінних, що є поняттям кластера (або k-ї теми). При цьому, якщо вдалося визначити максимальну семантичну близькість між термінами з множини $\mathbb{C}$ та наборів термінів з множини Ф, то елементу $y_к \in Y$ привласнюється значення еквівалентного терміна Ср,

Рішення подзадачі витягу ієрархічних зв'язків між поняттями ПрО засноване на ієрархічній агломеративної кластеризації. Вхідними даними підзадачі 3 є матриця Ф і вектор понять Y. Вихідними даними – матриця ієрархічних зв'язків понять Н та дендрограма.

Ієрархічна агломеративна кластеризація полягає в:

– обчисленні відстаней між темами-кластерами;

– кластеризації тем з урахуванням інформації про поняття.

Потім поєднуються дві найближчі по відстані dist теми (тобто відстань DIST між якими є мінімальним) t1 і tk; витягає відповідна до тем інформація з вектора понять Y і створюється новий кластер понять. Між новим кластером тем і іншими кластерами тем обчислюються відстані, і розмірність матриці відстаней зменшується на одиницю.

Об'єднання кластерів завершується, якщо виконується зупинний критерій або якщо всі кластери поєднуються в один. Мітки нових кластерів тем – понять визначаються шляхом об'єднання відповідних до цих тем поняттям.

На третьому етапі аналітичного методу формалізації ЛЕІ вирішується підзадача 3 – витяг ієрархічних зв'язків між поняттями. Вихідними даними

подзадачі 3 є вектор понять, який є результатом подзадачи 2. Між поняттями обчислюються відстані з обліком їх ймовірностей, і формується матриця попарних зв'язків понять (матриця відстаней). По отриманій матриці виконується ієрархічний кластерний аналіз, будується дендрограма – бінарне дерево, яке описує структуру вкладеності з ребрами (імовірності) між кластерами понять.

Структура пропонованого аналітичного методу формалізації ЛЕІ наведено на рисунку 2.1.

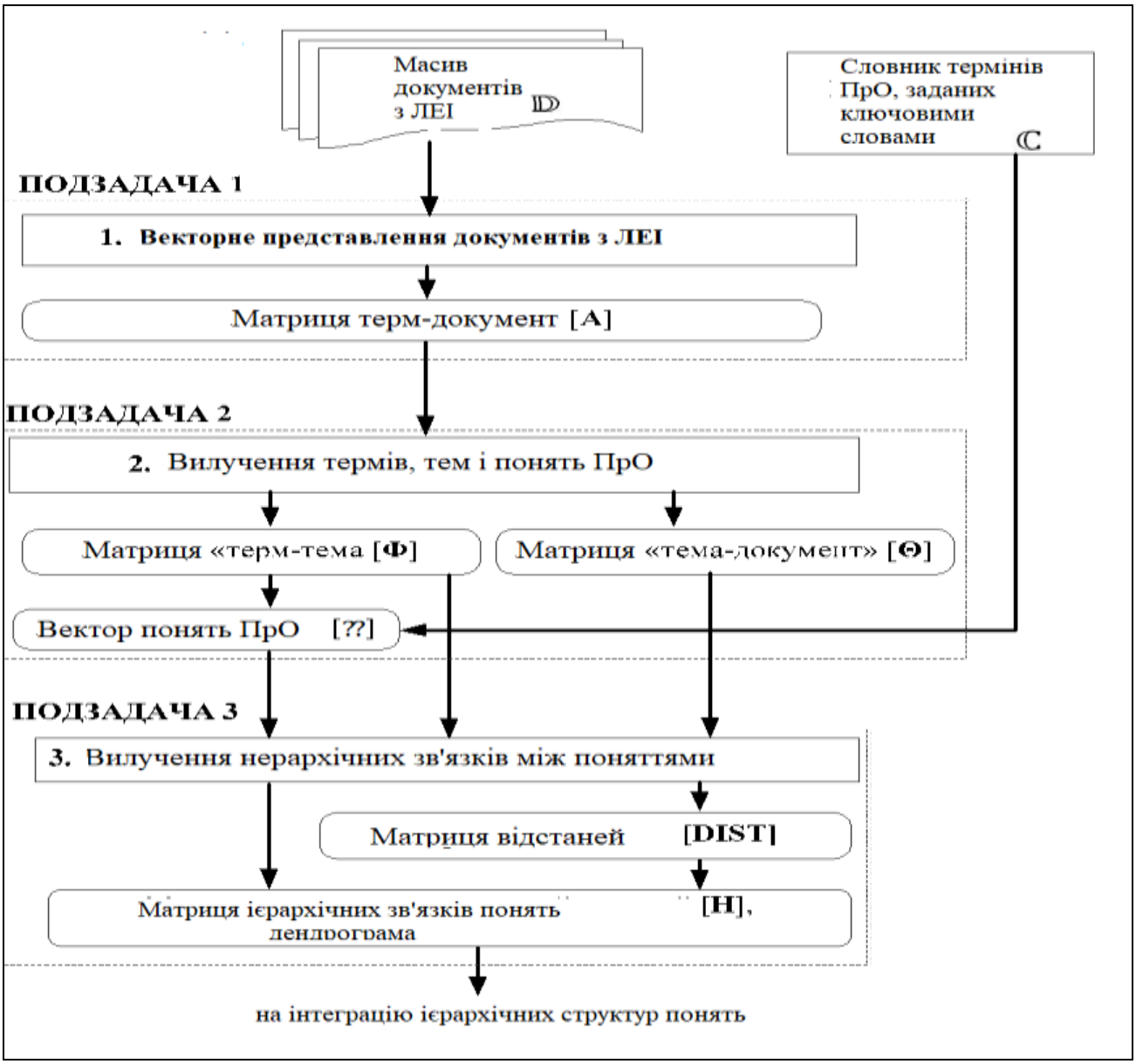

Рисунок 2.1 – Система розв'язуваних підзадач аналітичного методу формалізації лінгвістичної експертної інформації

Векторне представлення ЛЕІ (підзадача 1). Прийнявши гіпотезу «мішок термів», представлено документ d_i у вигляді вектор-стовпця довжиною, рівній довжині

словника N_W, елементом якого є ваговий коефіцієнт u_{ii} терма w_i у цьому документі, і відображення множину документів ЛЕІ $\mathbb{D}$ у числову двовимірну матрицю $\boldsymbol{D}_{i,j}$ (матриця терм-документ), рядки якої відповідають термам, стовпці – документам, що зустрічаються в усій множині документів $\mathbb{D}$, а елементом матриці є ваговий коефіцієнт u_{ii}:

$$\boldsymbol{D}_{j,i} = \begin{bmatrix} v_{11} & \cdots & v_{1i} & \cdots & v_{1N_D} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ v_{j1} & \cdots & v_{ji} & \vdots & v_{jN_D} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ v_{N_W1} & \cdots & v_{N_Wi} & \cdots & v_{N_WN_D} \end{bmatrix}, dim(\boldsymbol{D}_{j,i}) = N_W \times N_D, \tag{2.1}$$

де j = 1, …, N_w – номер рядка матриці $\boldsymbol{D}$, за номером терма w_j в словнику $\mathbb{W}$;

i = 1,.., N_D – номер стовпця матриці D, за номером документа d_I у $\mathbb{D}$.

u_{ij} – ваговий коефіцієнт j-го терма w_j для документа d_i:

$$u = tf(w, d), \tag{2.2}$$

де $tf(w, d)$ – частота терма w.

Частота терму дорівнює відношенню числа входжень терма w у документ d (wd) до загального числа термів цього документа тобто оцінюється важливість терма w у межах документа d:

$$tf(w, d) = \frac{n_{wd}}{N_{wd}}. \tag{2.3}$$

Приклад векторного представлення множини документів ЛЕІ для N_D = 3 показано на рис. 2.2.

Отримана терм-документна матриця D використовується для вирішення наступної підзадачі – витяг термів, тем і понять ПрО.

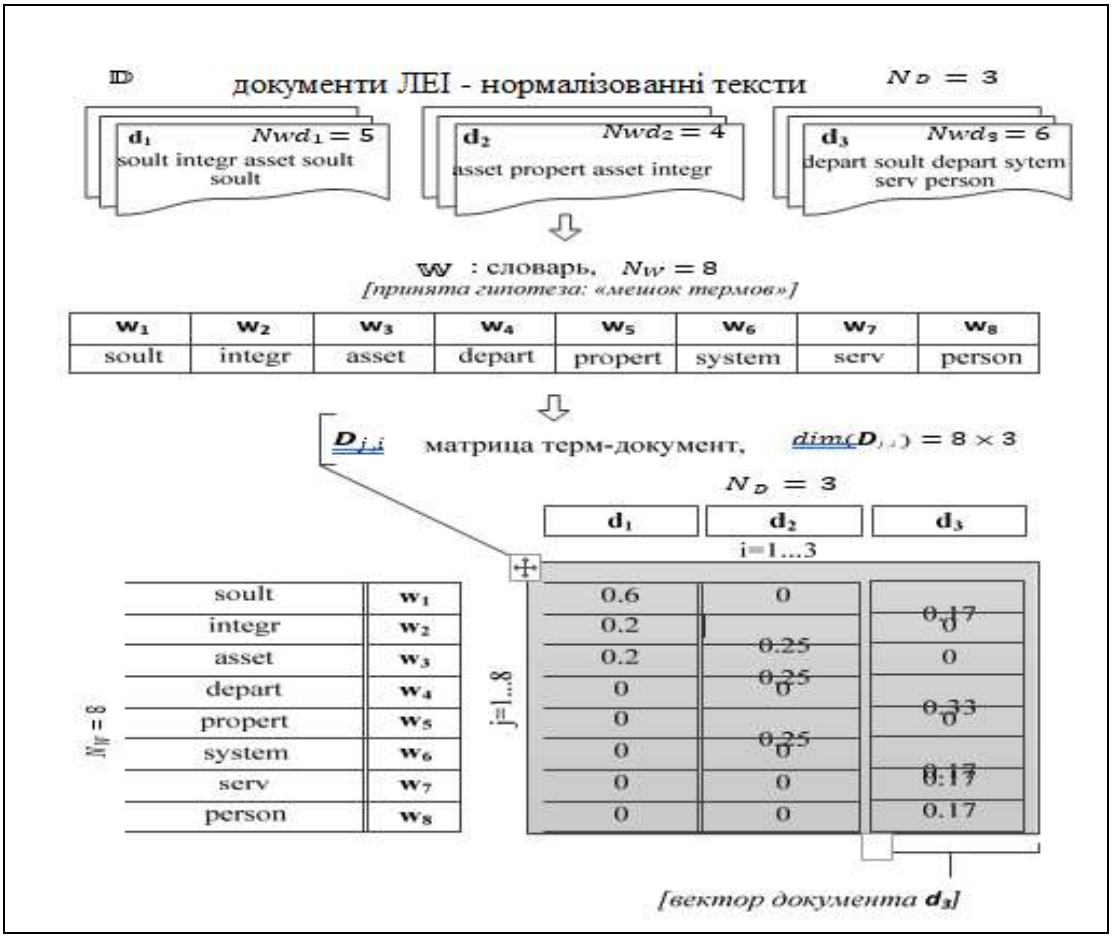

Рисунок 2.2 – Приклад векторного представлення множини документів ЛЕІ у вигляді матриці терм-документ

Рішення підзадач витягу термів, тем і понять засноване на побудові тематичної моделі множини документів ***D***.

Визначення матриці тем у документах і матриці термів у темах. Основною ідеєю ВТМ на основі ЛРД є те, що документи представляються сумішшю розподілів схованих тем t з множини можливих $\mathbb{T}$ $(t \in \mathbb{T})$, де кожна тема визначається імовірнісним розподілом на множина термів [12].

Представимо документи ЛЕІ як множину трійок (d, w, t), обраних випадково й незалежно з дискретного розподілу $p(d, w, t)$, заданого на кінцевій множині $\mathbb{D} \times \mathbb{W} \times \mathbb{T}$. ВТМ описує умовні розподіли термів у документах $p(w|d)$, виражаючи їх через умовні розподіли термів у темах $p(w|t)$і тем у документах $p(t|d)$:

$$p(w \mid d) = \sum_{t \in T} p(t \mid d)\, p(w \mid t). \tag{2.4}$$

Тематичним моделюванням називається рішення зворотного завдання в наступному формулюванні.

Множина документів $\mathbb{D}$ є відомим, причому кожний документ $d_j \in \mathbb{D}$ $(dim(d_j) = N_w d_j)$ розглядається як спостережувана випадкова незалежна вибірка термів $w_j \in \mathbb{W}$ $(dim(\mathbb{W}) = N_w)$, породжена деякою схованою підмножиною тем $\mathbb{T}$. Передбачається, що кожний документ d_j може відноситися до одної або декількох тем t_k з множини $\mathbb{T}$ ($t_k \in \mathbb{T}$, $k = 1,\ldots, N_T$, $N_T = \dim(\mathbb{T})$). Теми відрізняються друг від друга різною частотою вживання термів з $\mathbb{W}$.

Необхідно знайти: $\boldsymbol{\Theta}$ – матрицю розподілу тем на множині $\mathbb{D}$ і визначити, якою саме підмножиною тем породжений кожний документ; $\boldsymbol{\Phi}$ – матрицю розподілу термів, характерних для кожної теми.

У якості вихідної інформації тематична модель використовує матрицю $\boldsymbol{D}_{j,i}$ (2.1). Враховуючи, що умовні розподіли термів у документах p(w|d) можна знайти по спостережуваних частотах термів, обумовлених по формулі (2.3):

$$p(w|d) = \frac{n_{wd}}{N_{wd}}, \tag{2.5}$$

де n_{wd} – число входжень терма w в документ d;

N_{wd} – довжина документа d у термах w.

Матрицю терм-документ (2.1) в імовірнісному просторі представлено як:

$$\boldsymbol{D}_{j,i} = \begin{bmatrix} p(w_1|d_1) & \cdots & p(w_1|d_i) & \cdots & p(w_1|d_{N_D}) \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ p(w_j|d_1) & \cdots & p(w_j|d_i) & \vdots & p(w_j|d_{N_D}) \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ p(w_{N_W}|d_1) & \cdots & p(w_{N_W}|d_i) & \cdots & p(w_{N_W}|d_{N_D}) \end{bmatrix}, \tag{2.6}$$

Елементом гнізда матриці є імовірнісний розподіл термів (рядка) у документах (стовпці).

З метою одержання зв'язки відкритих і схованих параметрів, і згідно з теоремою про сингулярне розкладання [13], матриця (2.6) може бути розкладена на добуток двох матриць:

$$\boldsymbol{D} = \boldsymbol{\Phi} \cdot \boldsymbol{\Theta}, \tag{2.7}$$

де $\boldsymbol{\Phi}$ – двовимірна матриця імовірнісних розподілів термів за темами.

Матриця $\boldsymbol{\Theta}$ складається з N_D вектор-стовпців тем $\boldsymbol{\Theta}^{i}$.

Оцінка шуканих параметрів $\boldsymbol{\Phi}$ та $\boldsymbol{\Theta}$, враховуючи (2.4), може бути виконана за допомогою формули згорнутого семплювання Гибса [14].

Для знаходження вектора понять ПрО необхідно векторний простір знайдених тем у документах відобразити в простір термів. Оскільки термін – це слово або словосполучення, яке є назвою деякого поняття, після відображення тем у простір термів одержується поняття.

Застосуємо зовнішнє джерело знань – словник термів, відповідний до термінології предметної області. Словник термів ПрО являє собою множину $\mathbb{C}$, яке складається з N_C ($N_C = dim(\mathbb{C})$) векторів термів. Причому кожний термін $\boldsymbol{C}_p$ заданий кінцевою множиною ключових слів $c_l \in C_p$, $C_p = \{c_1, c_2, \ldots, c_{N_l}\}$. Число ключових слів N_L (для спрощення опису) узяте однаковим для кожного з термів C_p.

Використовуючи метод зворотної частоти, визначається вага ω кожного ключового слова c_l щодо множин $\mathbb{D}$ і $\mathbb{T}$ (для спрощення представлення індекси матриць і векторів не використовуються) [15]:

$$\omega = tfidf(\mathrm{c}, \mathbb{D}, \mathbb{T}) = tf(\mathrm{c}, \mathbb{D}) \cdot idf(\mathrm{c}, \mathbb{T}), \tag{2.8}$$

частота c ключового слова $tf(\mathrm{c}, \mathbb{D})$, дорівнює відношенню числа входжень ключового слова c (n_{cD}), обраного з терміна C, у множині $\mathbb{D}$ до загального числа термів цієї множини $\mathbb{D}$:

$$tf(\mathrm{c}, \mathbb{D}) = \frac{n_{cD}}{\sum_{i=1}^{N_D} Nwd_i}; \tag{2.9}$$

$tf(\mathrm{c}, \mathbb{T})$ – інверсна частота ключового слова c, з якої це ключове слово зустрічається в темах множини $\mathbb{T}$; обчислюється як відношення загального числа

тем t_k з множини $\mathbb{T}$ (N_T) до числа тем t_k з множини $\mathbb{T}$, у яких зустрічається ключове слово c ($N_T(c)$):

$$idf(\mathrm{c}, \mathbb{T}) = \lg\left(\frac{N_T}{N_T(c)}\right). \tag{2.10}$$

Використовуючи отримані ваги, відобразиться множина $\mathbb{C}$ у матрицю векторів ключових слів.

Одержано матрицю ваг ключових слів щодо множин $\mathbb{D}$ і $\mathbb{T}$:

$$\boldsymbol{\Omega}_{lp} = [\Omega_1 \quad \cdots \quad \Omega_{N_C}] = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1p} & \cdots & \omega_{1N_C} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \omega_{l1} & \cdots & \omega_{lp} & \vdots & \theta_{kN_C} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ \omega_{N_L1} & \cdots & \omega_{N_Lp} & \cdots & \theta_{N_LN_C} \end{bmatrix},$$
$$dim(\boldsymbol{\Omega}) = N_L \times N_C, \tag{2.11}$$

При формуванні вектора $\boldsymbol{\Phi}^{\kappa}$ вибирається N_L найбільш імовірних слів, що формують тему t_k, тобто для яких $p(w_j|t_k)$ відранжовані від максимального до мінімального значення у векторі-стовпці матриці $\boldsymbol{\Phi_{j,k}}$. При цьому із семантичної точки зору відсікання слів у темі, що мають меншу ймовірність, не впливає на зміст теми.

По отриманій мері близькості $\mathcal{S}^{\mathbb{T},\mathbb{C}}$ між векторами термів і тем для кожного терміна $\mathbf{C_p}$ вибирається тема t_k, для якої було отримано максимальне значення міри $\mathcal{S}^{\mathbb{T},\mathbb{C}}$ з терміном $\mathbf{C_p}$, тобто визначається максимальне значення міри близькості s_j теми t_k до терміна $\mathbf{C_p}$.

Будуються пари тема-термін (t_k, C_p) з позначенням відповідного міри близькості між ними (рис. 2.3).

Потім перевіряється еквівалентність пар. За фактом установлення еквівалентності термів і понять для кожної теми ставиться в мітка, відповідна до назви еквівалентного їй терміна ($C_{экв}$, абсолютна відповідність).

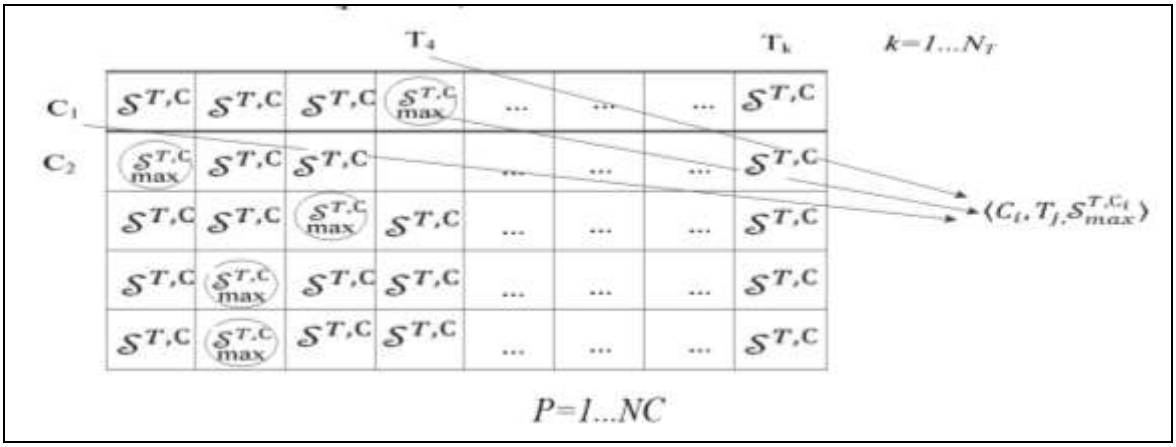

Рисунок 2.3 – Побудова пар термін-тема з максимальною мірою близькості

У випадку рівності заходів близькості одної теми (для якої вже встановлена абсолютна відповідність) із двома різними термінами $\mathbf{C_a}$ й $\mathbf{C_{p+1}}$, для наступної пари $\mathbf{C_{p+1}}$ і tk повторно виконується аналіз значень заходів близькості між цим терміном C_{p+1} і іншими темами за винятком теми, для якої абсолютна відповідність була встановлена. Вибирається максимальне значення міри близькості й виконується перевірка еквівалентності відповідно до граничного значення δ. За фактом установлення еквівалентності терміна й поняття для теми ставиться в мітка, відповідна до назви еквівалентного їй терміна ($C_{экв}$, часткова відповідність).

У випадку відсутності еквівалентності межу терміном і темою, тобто якщо $\mathcal{S}^{tk,Ca} < \delta$, створюється нове поняття. Із цією метою проводиться оцінка ймовірностей розподілу термів $\mathbf{W_j}$ у темі t_k, вектор значень яких отриманий по формулі (2.13). Вибирається терм із максимальним значенням імовірності p(w|t), і для цієї теми ставиться у відповідність мітка, відповідна до назви терма ($W_{екв}$, умовна відповідність).

2.3 Витяг ієрархічних зв'язків між поняттями

Виконується відображення вектора термів $\mathbf{C_p}$ і матриці розподілу термів **Φ** у вектор понять **Y.** Довжина вектора dim(Y) = Nт. Елементами вектора є строкова

змінна, що є поняттям кластера (або k-й теми). При цьому, якщо вдалося визначити максимальну семантичну близькість між термінами з множини $\mathbb{C}$ та наборів термів з множини $\mathbb{T}$, то елементу $\mathbf{y_к} \in \mathbf{Y}$ привласнюється значення еквівалентного терміна C_p, а якщо ні, то елементу $\mathbf{y_к} \in \mathbf{Y}$ привласнюється значення терма w_i, що має максимальне значення умовної ймовірності $p(w_j|t_k)$ для теми tk матриці **Ф.** Таким чином, вектор понять **Y** отриманий.

Рішення підзадачі витягу ієрархічних зв'язків між поняттями ПрО засноване на ієрархічній агломеративної кластеризації. Вхідними даними підзадачі 3 є матриця **Ф** і вектор понять **Y**. Вихідними даними – матриця ієрархічних зв'язків понять **H** та дендрограма. Завдання ієрархічного кластерного аналізу полягає в розбивці множини тем и понять на m непересічних кластерів.

Ієрархічна агломеративная кластеризація полягає в:

– обчисленні відстаней між темами-кластерами;

– кластеризації тем з урахуванням інформації про поняття.

Визначення відстані між темами. Кожна тема (спостережуваний об'єкт) характеризується j-ознаками (ознаками є ймовірності термів $\varphi_{jk} = p(w_j|t_k)$. Застосовується евклідова відстань для оцінки відстані між темами.

Після обчислення попарних відстаней між темами одержано квадратну й симетричну щодо головної діагоналі матрицю відстаней DIST:

$$\boldsymbol{DIST} = \begin{bmatrix} dist_{11} = 0 & dist_{12} & \cdots dist_{1k} \\ dist_{21} & dist_{22} = 0 & \cdots dist_{2k} \\ \vdots & \vdots & \ddots \ \vdots \\ dist_{k1} & dist_{k2} & \cdots \ 0 \end{bmatrix}, \dim(\boldsymbol{DIST}) = k \times k,\ k = \overline{1, N_T} \tag{2.12}$$

Кластеризація тем з урахуванням інформації про поняття. Спочатку кожна k - я тема $t_k \in \mathbb{T}$ розглядається як окремий кластер. Кожній темі у відповідність поставлене поняття з вектора понять **Y.** На першому кроці відстані між k темами-кластерами t_k визначаються по обраній мірі.

Потім поєднуються дві найближчі по відстані dist теми (тобто відстань DIST між якими є мінімальним) t_1 і t_k; витягає відповідна до тем інформація з вектора понять **Y** і створюється новий кластер понять. Між новим кластером тем і іншими

кластерами тем обчислюються відстані, і розмірність матриці відстаней ***DIST***, $\dim(\boldsymbol{DIST}) = k \times k$ зменшується на одиницю. На n-ом кроці процедура повторюється для матриці ***DIST***, $\dim(\boldsymbol{DIST}) = (k-n) \times (k-n)$ і т.д.

Об'єднання кластерів завершується, якщо виконується зупинний критерій або якщо всі кластери поєднуються в один (див. рис. 2.4). Мітки нових кластерів тем – понять визначаються шляхом об'єднання відповідних до цих тем поняттям.

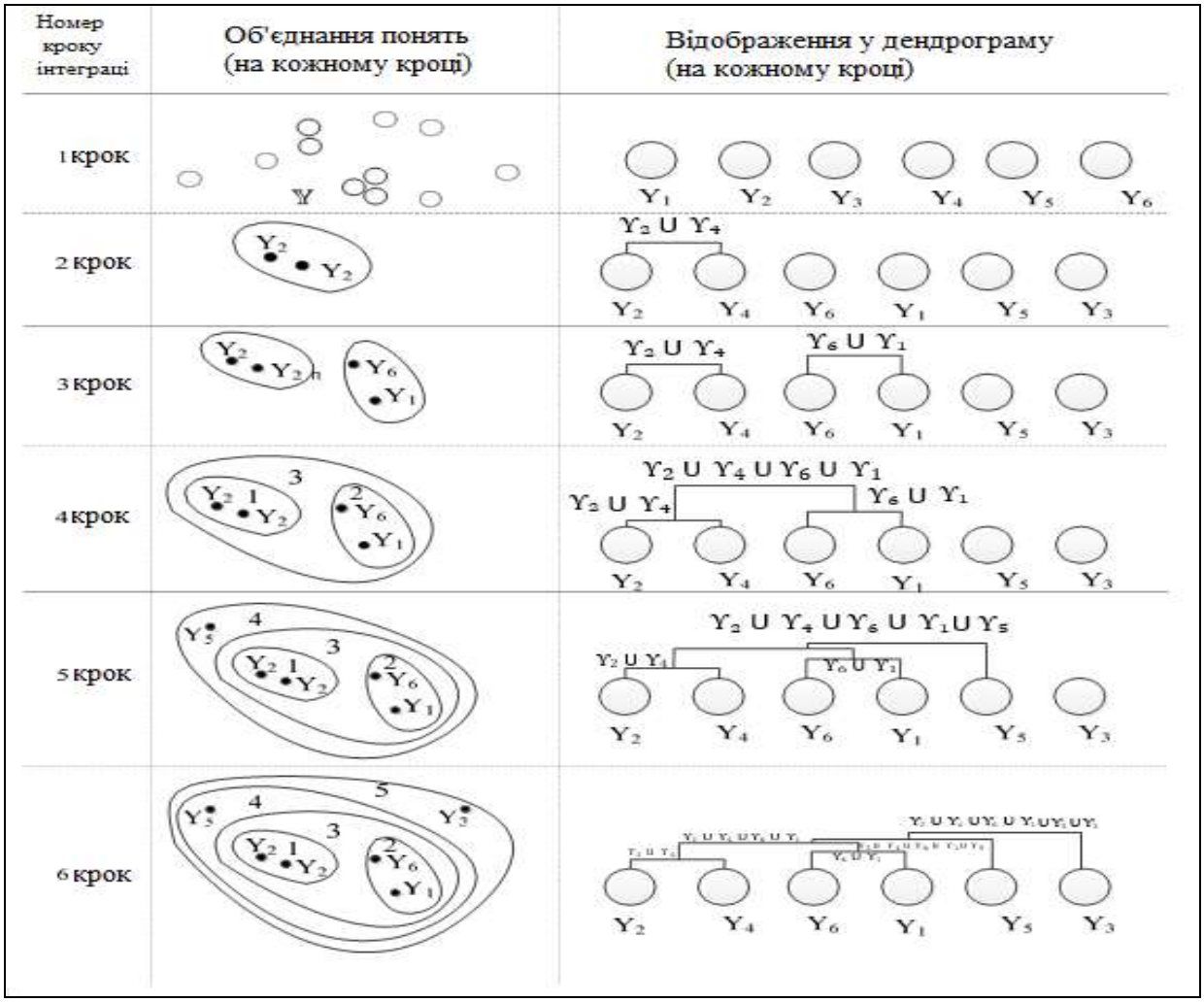

Рисунок 2.4 – Схема об'єднання понять у кластери

Нехай спочатку було 6 тем і кожній темі у відповідність було поставлене поняття: $t_k \rightarrow Yk$. На першому кроці кожне спостереження представляє один кластер (тобто є шість кластерів), на другому кроці відбувається об'єднання Y2 і Y4 та утворюється новий кластер з, міткою Y2 U Y4. На третьому кроці поєднуються кластери Y6 і Y1 з міткою нового кластера Y6 і Y1. Процес триває доти, поки всі спостереження не об'єднаються в один кластер.

Відстані між кластерами, коли необхідно зв'язати кілька тем і понять, визначаються за певним правилом зв'язку. Якість рішення в ієрархічній кластеризації залежить від способу зв'язування об'єктів у кластері й алгоритму їх об'єднання в нові. Традиційні методи зв'язування мають ряд недоліків. Наприклад, основним недоліком правила одиночному зв'язку є «жадібний» характер, що приводить до одержання локально оптимальних кластерів, крім того застосування цього правила не дозволяє розділяти пересічні кластери [18]. Недоліком правил середнього й повного зв'язування є висока обчислювальна трудомісткість, що не дозволяє використовувати їх для обробки більших масивів даних (рис. 2.5).

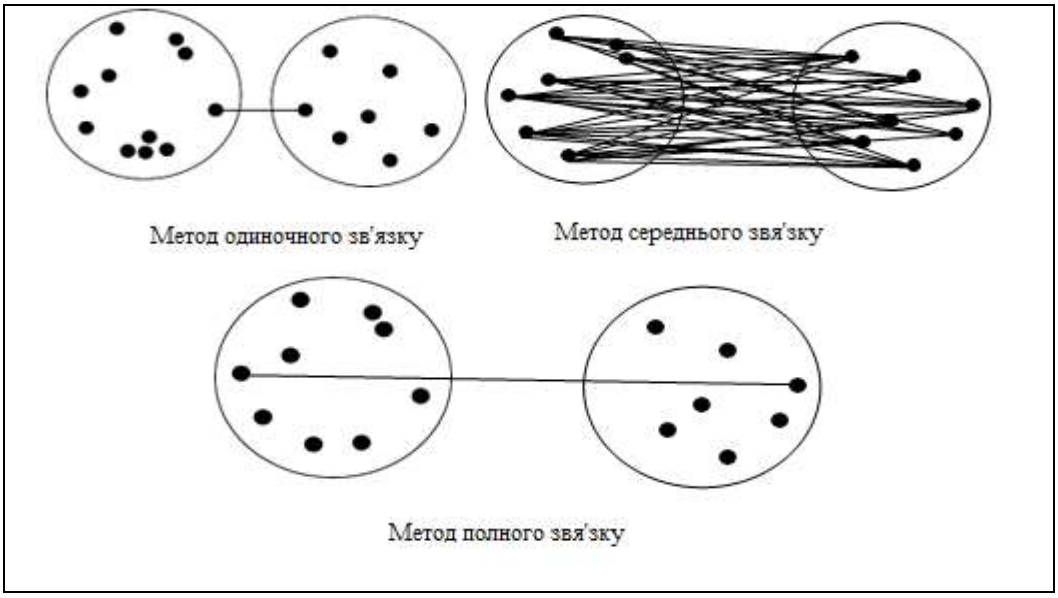

Рисунок 2.5 – Приклади правил зв'язування об'єктів у кластері

Для усунення виявлених недоліків у роботі пропонується два правила зв'язку кластерів, що полягають із декількох понять:

– середньому зв'язку по мінімальній відстані між об'єктами;

– середньому зв'язку по максимальній відстані між об'єктами.

Перше правило. Функція середнього зв'язку по мінімальній відстані визначається як середня відстань між найближчими парами тем, що входять у кластери.

На рисунку 2.6 показаний механізм обчислення зв'язки для об'єднання кластерів понять по методу середнього зв'язку по мінімальній відстані при q=3.

Друге правило – функція середнього зв'язку по максимальній відстані визначається як середня відстань між ke найбільш вилученими парами об'єктів γ^x і γ^y, де ($\gamma^x \in Y_x, \gamma^y \in Y_y$:

При цьому аргумент максимізації – значення аргументу, при якому дане вираження досягає максимуму. На рис. 2.7 показаний механізм обчислення зв'язки для об'єднання кластерів понять по методу середнього зв'язку по максимальній відстані при q=3.

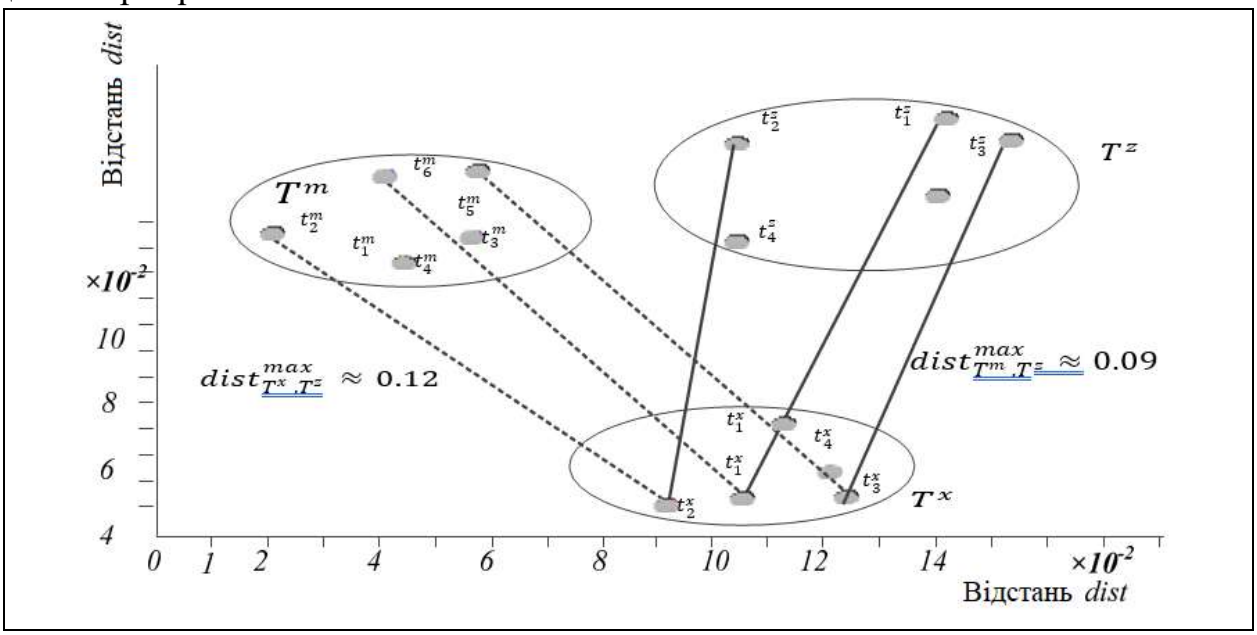

Рисунок 2.6 – Правило середнього зв'язку по мінімальній відстані

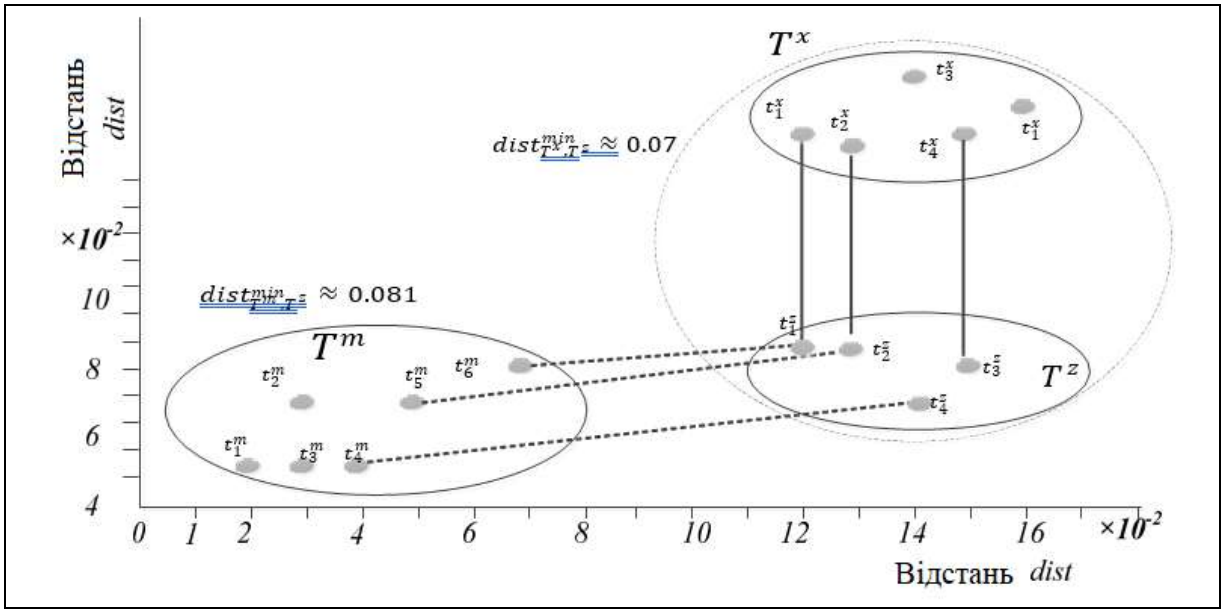

Рисунок 2.7 – Правило середнього зв'язку по максимальній відстані

Результатом об'єднання тем у кластери є матриця ієрархічних зв'язків – H розмірністю (k - 1) × 5, де k – число первісно заданих об'єктів спостереження (k) У першому й другому стовпцях матриці втримуються індекси попарно згрупованих кластерів (таким чином, формується бінарне дерево). Листами бінарного дерева є спочатку задані теми t_k. Кожному знову створюваному кластеру рядка H(I,:), привласнюється індекс $k + I$. Елементи матриці H(I,1) і H(I, 2) містять індекси двох кластерів, які утворюють кластер $k + I$ Кластери, що випливають після k - I є внутрішніми вузлами дендрограми. У третьому й четвертому стовпцях утримуються поняття, відповідні до індексів тем. У п'ятому стовпці матриці H зберігаються відстані між двома кластерами, які поєднуються у відповідному рядку H(I).

По отриманій матриці ієрархічних зв'язків будується дендрограма, що представляє бінарне дерево, яке описує структуру вкладеності з ребрами між вкладеними кластерами тем (понять). Дендрограма показує ступінь близькості окремих об'єктів і кластерів, а також наочно демонструє в графічному виді послідовність їх об'єднання або поділи. Кількість рівнів дендрограми відповідає числу кроків злиття або поділи кластерів.

Отримані матриця ієрархічних зв'язків **H** з обліком тематичної імовірнісної близькості між поняттями, а також дендрограма, які можуть використовуватися як для інтеграції ЛЕІ різних колекцій текстів, так і для побудови онтології.

3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНЬ

3.1 Аналітичний метод інтеграції ієрархічних структур

Інтеграція ІСП дозволяє поєднувати формалізовані представлення різних колекцій текстів, що перебувають у різних джерелах. Інтеграція формалізованих даних у вигляді ІСП виконується на семантичному рівні й полягає в побудові єдиної ІСП для декількох колекцій текстів [23].

Вхідними даними аналітичного методу інтеграції ІСП є результати, отримані за допомогою аналітичного методу формалізації ЛЕІ, застосованого для різних колекцій текстів, отриманих у різний час із джерел A и B, але приналежних до одної ПрО: матриці терм-тема Φ_A і Φ_B, терм-документ Θ_A і Θ_B і ієрархічних зв'язків понять H_A і H_B За вхідним даними необхідно знайти подібність двох ієрархічних структур H_A і H_B і створити нову ієрархічну структуру H_U, що поєднує, що й погодить семантичні представлення вихідних колекцій текстів.

В основу аналітичного методу інтеграції ІСП покладений підхід обчислення комплексного міри семантичної близькості, елементи якої обчислюються на рівні термів, рівні понять і структурному рівнях. Для згортки елементів комплексного міри близькості застосовується адитивна функція з урахуванням важливості кожного елемента, вираженої ваговим коефіцієнтом [4]. Для одержання оптимальних значень вагових коефіцієнтів, застосовується еволюційний алгоритм – метод роя часток. Структура пропонованого аналітичного методу інтеграції ієрархічних структур понять декількох колекцій текстів показано на рис. 3.1.

Крок 1 – пошук локальних матриць семантичної близькості: матриці близькості термів, матриці близькості понять і матриці близькості таксономій.

Оцінка семантичної близькості на рівні термів виконується між кожними вектор-стовпцями відповідно, отриманих із джерел A и B. Оцінка коефіцієнтів семантичної близькості на основі косинусної міри обчислюється як міра подібності між двома векторами простору:

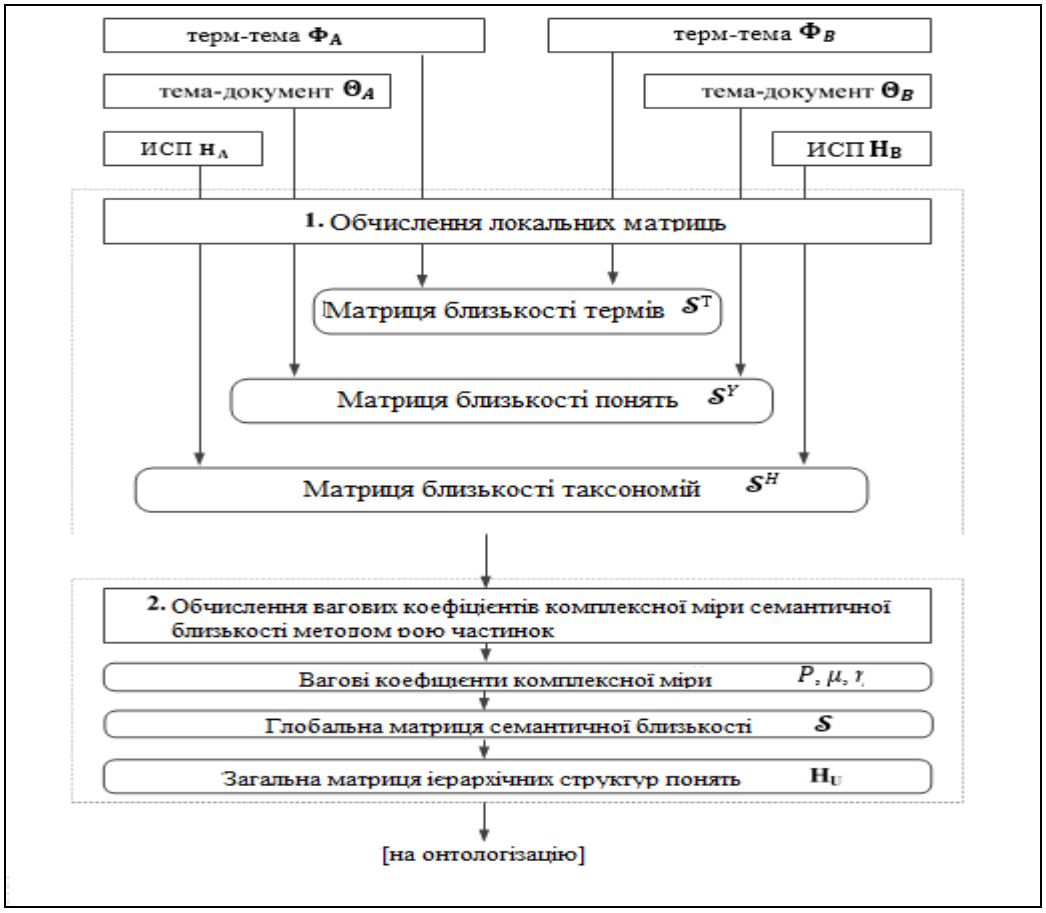

Рисунок 3.1 – Структура аналітичного методу інтеграції ієрархічних структур понять декількох колекцій текстів

Оцінка семантичної близькості на рівні понять виконується між кожними вектор-стовпцями матриць.

Обчисливши коефіцієнти семантичної близькості на рівні понять, одержано матрицю семантичної близькості понять:

$$\boldsymbol{S}^{\boldsymbol{T}}_{\boldsymbol{a,b}} = \begin{bmatrix} S^T_{11} & \cdots & S^T_{1b} & \cdots & S^T_{1N_T^{(B)}} \\ \vdots & \ddots & & \vdots \; \vdots & \vdots \\ S^T_{a1} & \cdots & S^T_{ab} & \vdots & S^T_{aN_T^{(B)}} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ S^T_{N_T^{(A)}1} & \cdots & S^T_{N_T^{(A)}b} & \cdots & S^T_{N_T^{(A)}N_T^{(B)}} \end{bmatrix}, dim(\boldsymbol{S}^{\boldsymbol{T}}_{\boldsymbol{a,b}}) = N_T^{(A)} \times N_T^{(B)}, \tag{3.1}$$

де $q = 1, N_D$ – номер стовпця матриці $\boldsymbol{\Theta_B}$, обумовлений порядковим номером d_i у множині $\mathbb{D}$ джерела B;

$\mathcal{S}^{T_{\text{eq}}}$ – коефіцієнт семантичної близькості на рівні термів, обчислений по формулі.

Оцінка семантичної близькості на таксономічному рівні. Близькість двох понять оцінюється по положенню вершин, відповідних до цих понять в ієрархічних структурах. Міра близькості таксономічної ієрархії заснована на довжині найкоротшого шляху, вимірюваного числом вершин (або ребер) у шляху між двома відповідними вершинами таксономії, з урахуванням глибини таксономічної ієрархії – чим менше довжина шляху між вершинами, тем вони ближче [20]. На рисунку 3.2 показаний фрагмент таксономічного дерева для прикладу, описаного на рисунку 2.4, крок 5.

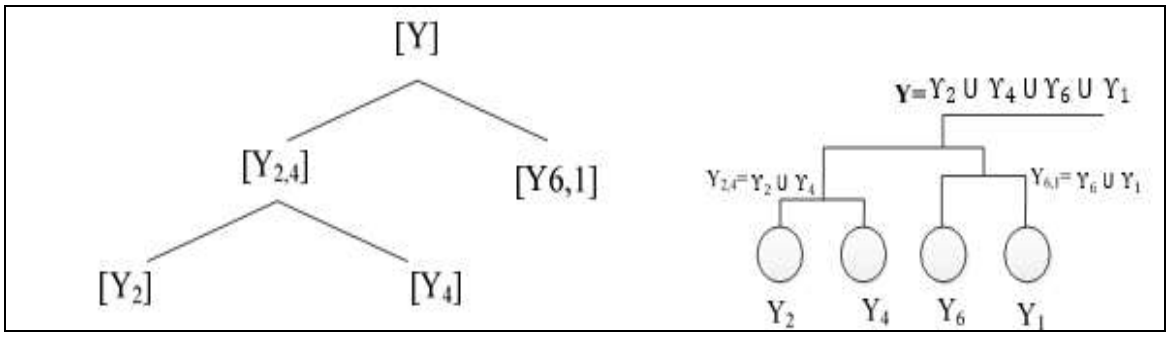

Рисунок 3.2 – Визначення довжини найкоротшого шляху

Довжина шляху між вершинами $Y_{2,4}$ і $Y_{6,1}$ дорівнює 3 (число вершин). Довжина шляху між вершинами Y_2 і Y_4 теж 3. Однак інтуїтивно близькість першої пари менше, ніж другий, яка належить більш низькому рівню й розділяє більше інформації, ніж перша пара. Чим нижче рівень пари вершин, тем вони семантично ближче в порівнянні з парами більш високого рівня. У міру спуска рівень семантичної деталізації збільшується.

3.2 Модифікація алгоритму обчислення міри семантичної близькості

Вагові коефіцієнти, які регулюють процес обчислення семантичної близькості понять:

$$S(\gamma_a^{(A)}, \gamma_b^{(B)}) = \ell \cdot S^T(\gamma_a^{(A)}, \gamma_b^{(B)}) + \mu \cdot S^C(\gamma_a^{(A)}, \gamma_b^{(B)}) + \eta \cdot S^H(\gamma_a^{(A)}, \gamma_b^{(B)}), \quad (3.3)$$

де $S \in [0;1]$, $S = 1$, якщо поняття еквіваленти; $S = 0$, якщо поняття різні.

Обчислення вагових коефіцієнтів комплексної міри семантичної близькості понять двох ієрархічних структур.

Завдання обчислення вагових коефіцієнтів комплексної міри семантичної близькості відноситься до групи завдань оптимізації з обмеженнями. Застосовано еволюційний підхід для знаходження вагових коефіцієнтів комплексної міри семантичної близькості. У якості загальної структури алгоритму використовувався модифікований (дискретний) метод роя часток [26]. У методі роя часток у ролі часток виступають вагові коефіцієнти семантичної близькості

В такому разі завдання обчислення вагових коефіцієнтів комплексної міри семантичної близькості виглядає як завдання багатопараметричної оптимізації:

$$\min_{\vec{x}} f(\vec{x}),$$
$$\vec{x} = \ell, \mu, \eta \in F \subseteq S, \ell, \mu, \eta \in [0;1], \ell + \mu + \eta = 1 \quad (3.3)$$

де $\vec{x}$– вектор рішень, називається припустимим рішенням F,

S – уся область пошуку,

f: $R^n \to R$ при $k(k \geq 1)$ цільова функція.

Необхідно знайти $\vec{x} \in F$ такий, що $f(\vec{x}') \geq f(\vec{x}) \forall \vec{x} \in F$.

У якості цільової функції, залежно від наявності, що є в наявності, можуть використовуватися різні критерії. Наприклад, якщо задане еталонне відображення, можна в якості критеріїв задати параметри точності й повноти зображення (докладний опис методу стосовно до завдання відображення онтологій із застосуванням роя часток наведене в [23]). Якщо еталонне відображення є невідомим, у якості критерію можна задати відстань близькості між елементами двох понять ІСП (γ_A й γ_B), наприклад, відстань Евкліда [9]:

Початкова популяція часток задається випадково згенерованим набором значень із обмеженнями $P, \mu, \eta \in [0; 1], P + \mu + \eta = 1$. Популяція називається роєм і вона складається з m чисел підходящих рішень або часток. Кожна частка має n позицій, що містять n вагових коефіцієнтів, що відповідають n різним мірам подоби.

Нехай f – цільова функція, яку потрібно мінімізувати, N-кількість часток у рої, кожній з яких зіставлена координата $\mathbf{x}_i \in \mathbb{R}^n$ у просторі рішень і швидкість $v_i \in \mathbb{R}^n$.Краще з відомих положень частки pbest, найкращий відомий стан роя в цілому gbest. Створення роя наступного покоління виконується шляхом оцінки положення й швидкості частки. Кожне гніздо або позиція являє собою вагу (нормалізоване значення гнізда) щодо міри подоби. Гнізда усередині частки містять значення від 0 до 1, а швидкостям кожного гена задані нульові значення. Використовуючи інформацію, отриману на попередній стадії, положення кожної частки й швидкість кожного кластера оновлюється. Кожна частка відслідковує кращу позицію, якої вона досягла (pbest). З погляду багатокритеріального підходу, вибирається краща позиція, для якої пристосованість цієї частки домінує над іншими пристосуваннями. Найкраще положення серед усіх часток називається глобальним кращим.

Значення швидкості й координати представлені у вигляді векторних значень і містять множину компонентів, а не просто одне скалярне значення. Кількість ітерацій і часток у рої вказують на тривалість роботи алгоритму, а, отже, і на якість одержуваних результатів; значення констант визначаються експериментально виходячи з поставленого завдання.

Далі обчислюється значення цільової функції кожної частки. На основі отриманих даних проводиться вибір часток, які дають найменше значення цільової функції.

Критерієм зупинки є досягнення заданого числа ітерацій або необхідного значення цільової функції. Пристосування частки виконується доти, поки не буде досягнутий критерій зупинки для переходу на наступну ітерацію.

Результат фіксується у вигляді популяції часток – вектора значень вагових коефіцієнтів семантичної близькості. Виконується підсумкове обчислення

комплексного міри семантичної близькості. Результатом етапу є глобальна матриця коефіцієнтів семантичної близькості.

Отримана глобальна матриця коефіцієнтів семантичної близькості використовується для побудови підсумкової ІСП. Над поняттями ІСП джерел А и В виконуються наступні операції:

– якщо $S > \delta$, то поняття Л джерела А и джерела В є еквівалентними – для них виконується операція об'єднання (правило 2);

– якщо поняття ІСП джерела А відповідає декільком поняттям ІСП джерела В (отримані однакові коефіцієнти заходів семантичної близькості, що задовольняють умові $S > \delta$), – виконується об'єднання понять результуючої ІСП на структурному рівні. У такому випадку поняття ІСП джерела В представляються як підклас поняття джерела А в результуючої ІСП (правило 3);

– якщо $S < \delta$, то поняття ІСП джерела А и джерела В не є еквівалентними – виконується операція додавання поняття ІСП джерела В у результуючу ІСП (правило 4).

Таким чином, за результатами виконання правил 1-4 формується підсумкова матриця U.

Таким чином, аналітичний метод формалізації ЛЕІ, отриманої з колекції текстів заданої предметної області, у вигляді ієрархічної структури понять, відрізняється від відомих обчисленням схованих закономірностей у тексті (за рахунок використання підходу імовірнісно-тематичного моделювання), кластеризації понять і відображенням їх ієрархічних зв'язків у вигляді дендрограми з обліком тематичної імовірнісної близькості між поняттями. Метод тематичного моделювання дозволяє без участі експерта обчислити мітки тематичних кластерів – мітки понять. Метод ієрархічної агломеративної кластеризації дозволяє зменшити обчислювальне навантаження при обчисленні відстаней між кластерами завдяки застосуванню нових правил середньому зв'язку по мінімальному/максимальній відстані. Комплексування підходів тематичного моделювання й ієрархічної кластеризації дозволяє побудувати ієрархічну структуру понять предметної області (дендрограмму) відмінної від відомих тим,

що крім ієрархічних зв'язків ураховуються зв'язки імовірнісної тематичної близькості.

В основу методу покладений алгоритм обчислення комплексної міри семантичної близькості, елементи якої обчислюються на рівні термів, рівні понять і структурному рівнях.

4 ОПИС РОЗРОБЛЕНОГО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

4.1 Структура алгоритму

Розроблені алгоритми формалізації ЛЕІ, що отримана з колекції текстів заданої предметної області і інтеграції ієрархічних структур понять декількох колекцій текстів, що належать заданої предметної області можуть мати прикладне застосування як математичне забезпечення системи інтелектуальної обробки текстової інформації. Система інтелектуальної обробки текстової інформації повинна складатися з інформаційного, алгоритмічного й програмного забезпечення. Структура побудови – модульна, нарощувана.

Вихідними даними алгоритму є: текстові документи, уміст яких належить одній предметній області, гіперпараметри розподілу Дирихлє, число тем, рівень зрізу дендрограми, параметри МРЧ.

Загальна структура алгоритму.

Крок 1 – формування модулю формалізації ЛЕІ:

– побудувати навчальну й контрольну вибірки;

– побудувати модель «мішок термів»;

– обчислити частоти зустрічання термів у документах;

– навчити модель витягу термів і тем (навчальна вибірка);

– настроїти параметри моделі;

– застосувати навчену модель витягу термів і тем (контр, вибірка).

Крок 2 – побудувати матрицю ключових слів і термів у документах і темах.

Крок 3 – обчислити міру близькості між векторами тем і термів:

– ранжувати коефіцієнти близькості векторів тем для термів;

– сформувати вектор понять за правилами еквівалентності;

– обчислити відстані між темами.

Крок 4: Задати метод об'єднання кластерів і виконати кластеризацию понять.

Крок 5 – Побудувати дендрограмму.

Крок 6 – виконати алгоритм «Модуль інтеграції ІСП (*A* і *B*)»:

– обчислити міру близькості термів;

– обчислити міру близькості понять;

– обчислити міру близькості таксономії;.

Крок 7 – ініціалізувати параметри: задати цільову функцію і виконати МРЧ.

Крок 8: обчислити комплексну міру семантичної близькості.

Крок 9 (фінальний): сформувати підсумкову ІСП та побудувати підсумкову дендрограмму.

В результаті отримано «Масив документів ЛЕІ», що має розмірність ND та містить:

– матриця « темаудокумент» [ПРО];

– матриця «терм-тема» [Ф];

– вектор понять Про [Y];

– матриця ієрархічних зв'язків понять [Н];

– дендрограмма.

Структурно алгоритм складається з наступних компонентів:

– модуля попередньої обробки колекції текстів;

– модуля формалізації ЛЕІ.

– модуля інтеграції ІСП.

Вихідними даними алгоритму є матриці розподілу тем по $\boldsymbol{\Theta} = (\theta ki)$, термів по документах $\boldsymbol{\Phi} = (\varphi jk)$, вектор понять $\boldsymbol{\Upsilon}$ ієрархічна структура понять Про у вигляді матриці Н и дендрограми.

Алгоритм дозволяє створити процедури одержання ЛЕІ з колекції документів, формалізації отриманої ЛЕІ у вигляді ієрархічних структур понять і інтеграції ієрархічних структур понять декількох колекцій текстів одні Про.

Процедура алгоритму реалізована із застосуванням функцій пакетів [31]:

– «Text analytics toolbox» – функції попередньої обробки, аналізу й моделювання текстових даних;

– «Statistics and machine learning toolbox» – функції статистичної обробки, кластерного аналізу;

– «Global optimization toolbox» – функції рішення завдань оптимізації функцій.

Розроблений алгоритм і його процедура може використовуватися як для рішення приватних завдань інтелектуальної обробки текстової інформації (витяг тематичних кластерів і тем з колекцій текстів, кластеризація, візуалізація змісту колекції текстів, інтеграція ієрархічних структур понять – для проведення контент-аналізу), так і для побудови простої онтології колекції текстів, заданої у вигляді ієрархічної структури понять предметної області.

Приведемо деталі процедури інтелектуальної обробки лінгвістичної експертної інформації для формалізації текстової інформації.

4.2 Модуль попередньої обробки колекції текстів

Модуль попередньої обробки колекції текстів виконує попередню обробку колекції текстів у форматах pdf, txt, doc(x), html, яка полягає в завантаженні текстів, видаленні з масиву документів зайвих цифр, стоп-слів, тегів, коротких і довгих слів, приведенні слів до одному регістру, а також лінгвістичній обробці КТ, реалізованої на морфемному й лексичному рівнях.

Функції виконуються послідовно: вихідні дані одної функції є вхідними даними для наступної функції.

Вихідними даними модуля попередньої обробки текстової інформації є масив документів ЛЕІ – множина D, що полягає з ND документів, тип даних string array.

Склад функцій модуля попередньої обробки колекції текстів наведено в таблиці 4.1.

Таблиця 4.1 – Функції модуля попередньої обробки КТ

Номер кроку	Функція	Назва функції	Опис
1	`extractfileText`	завантаження текстових	зчитує текст із форматів pdf, doc(x), txt, html на згадку
2	`erasepunctuation`	видалення розділових знаків з тексту	очищення тексту від непотрібних символів
3	`lower`	приведення символів до нижнього регістру	перетворення символів верхнього регістру до символів нижнього регістру
4	`tokenizeddocume`	виділення лексем	розбивка тексту на термін
5	`stopwords`	фільтрація стоп-слів	задає список слів
6	`removestopWords`	видалення стоп-слів	видаляє слова, задані в списку стоп-слів
7	`removeshortwords`	видалення коротких слів	видалення слів, які складаються з невеликого числа символів
8	`removelongwords`	видалення довгих слів	видалення довгих слів (більш 14 символів)

4.3 Модуль формалізації лінгвістичної експертної інформації

Вхідними даними модуля формалізації ЛЕІ є масив документів ЛЕІ – D, отриманий з модуля попередньої обробки колекції текстів.

Модуль формалізації ЛЕІ формує навчальну й контрольну вибірки за вхідним даними – масиву документів ЛЕІ, виконує перетворення документів ЛЕІ у векторне представлення, виділення термів, тем і понять ПрО, побудова ієрархічної структури понять ПрО.

Послідовність дій, виконуваних для формалізації ЛЕІ.

Крок 1. Побудувати навчальну й контрольну вибірки – функція cvpartition. Дія дозволяє розбити випадковим образом масив документів ЛЕІ D на навчальну й контрольну вибірки. У результаті буде отримано два масиви документів ЛЕІ (навчальна вибірка складається з 80-90% документів, інші документи ЛЕІ становлять контрольну вибірку).

Крок 2. Побудувати модель «мішок термів» – функція bagofwords. Дія виконується для двох масивів (навчального й контрольного) окремо. Результатом дії є:

– словник термів (Vocabulary = W = $\{w_{lt}\ldots,w_j,\ldots,w_{nw}\}$, тип даних string vector;

– число термів (Numwords = Nw), тип integer;

– число документів ЛЕІ (Numdocuments=ND) у масиві ЛЕІ, тип даних integer;

– число термів у кожному документі масиву ЛЕІ (Counts = Nwdt), тип даних string array.

Крок 3. Обчислити частоти зустрічання термів у документах –функція tf (2.3). Результатом дії є матриця терм-документ (2.1), тип даних string array;

Крок 4. Навчити модель витягу термів і тем (навчальна вибірка) – функція fitlda. Вхідними даними є: матриця терм-документ $D_{j,i}$, отримана для навчального підмножини; гіперпараметри α и β розподілу Дирихлє.

Для параметрів Дирихлє α и β виконуються наступні умови [13]:

якщо $\alpha_t = 1$ для всіх t, то розподіл Дирихлє переходить у рівномірне; чим більше α_0, тем менше дисперсія, і тем сильніше вектори тем концентруються навколо вектора математичного очікування. Чим менше α_0, тем сильніше значення θ_{ki} концентруються навколо нуля, тем сильніше різняться документи θ_i;

– параметр β визначає ступінь розрідженості векторів φ_{jk} матриці Ф, породжуваних розподілом Дирихлє – чим менше β_0, тем сильніше різняться теми φ_k.

Оцінку параметрів $\Phi = \varphi_{jk}$ і $\Theta = \theta_{ki}$ виконати з застосуванням алгоритму семплування Гіббса, задавши аргумент функції fitlda: solver = cvbo. Вихідними даними функції `fitlda` є:

– число тем у колекції (Numtopics = N_T), тип даних positive integer;

– концентрація тем (Topicconcentration), тип даних positive scalar;

– концентрація слів (Wordconcentration), тип даних nonnegative scalar;

– імовірності тем у документах (Corpustopicprob), тип даних, vector;

– імовірності тем у документі ($DocumentTopicProb = \theta_{ki}$) тип даних matrix;

– імовірності термів у темах ($TopicWordProb$=φ_{jk}), тип даних matrix, модель Idamodel.

Крок 5. Налаштування параметрів моделі – функція logp. Дія дозволяє виконати пошук оптимального значення параметра k – кількості схованих тем, по заданих параметрах розподілу Дирихлє. Налаштування параметрів виконується шляхом обчислення внутрішніх критеріїв оцінки якості моделі `Idamodel`.

Для вибору оптимальної кількості тем (T) ітеративно виконуються обчислення перплексії моделі для різного числа тем з деяким кроком (наприклад, 10, 20, 40, 80, і 160). Одержувані на кожному кроці значення перплексії по контрольній вибірці рівняються з мінімальним значенням. Модель із мінімальним значенням перплексії є оптимальною.

Результатом дії є навчена модель `ldaModel` , число тем, при якому досягається оптимальне значення перплексії ($k = N_T$).

Крок 6. Виконати перевірку моделі витягу термів і тем з документів ЛЕІ – застосувати функцію $fitlda$ до матриці терм-документ, отриманої для контрольної вибірки масиву документів ЛЕІ, задавши число тем.

Результатом кроку 6 є модель із мінімальним значенням перплексії. По оптимальнім числу тем формуються підсумкові розподіли тем по документах і термів по темах.

Крок 7. Побудувати матрицю ключових слів термів у документах $\mathbb{D}$ і темах $\mathbb{T}$ функція term-Dt. Дія виконується шляхом підрахунку частот зустрічання ключових слів у словнику термів з урахуванням інверсної частоти ключових слів, з якої це ключове слово зустрічається в темах множина. Результатом дії є матриця ваг ключових слів щодо множини.

Крок 8. Обчислити косинусну міру близькості між векторами тем і термів – функція *sem-cos*. Дія виконується шляхом попарного обчислення косинусної міри близькості між кожним вектором $\Phi^к$ (до = 1,NT) і кожним вектором $\mathbf{\Omega}^p$ (a = l,Nc). Результатом дії є матриця коефіцієнтів близькості векторів тем і термів, тип даних string array.

Крок 9. Ранжувати коефіцієнти близькості векторів тем і термів для кожного терміна – функція `rang`. Результатом дії є матриця з відсортованими по спаданню коефіцієнтами косинусної близькості векторів тем і термів, тип даних string array.

Крок 10. Сформувати вектор понять за правилами еквівалентності (правило 1) – функція gen-ecv. Результатом дії є вектор понять Y.

Крок 11. Обчислити відстані між векторами тем $\Phi^k_x = (x_1 ... x_j)$ і $\Phi^k_y = (y_1, ...y_j)$ – функція `pdist`. Аргументом функції pdist є Distance вибір, що визначає, метрики, по якій обчислюються відстані. Функція pdist повертає парні відстані між темами згідно (2.19). Результатом дії є матриця відстаней.

Крок 12. Задати метод об'єднання кластерів – функція linkage (Y), де Y – є аргументом функції вектор, що представляє собою, відстань між темами, що вертаються функцією pdist. Строкова змінна method, що задає правило об'єднання кластерів, ухвалює одне зі значень: single (алгоритм одиночному зв'язку), complete (алгоритм повного зв'язку); average (алгоритм середнього зв'язку) або правила середнього зв'язку по мінімальному й максимальному відстанях (2.21), (2.22), відповідно.

Крок 13. Виконати кластеризацію понять – функція cluster. Аргументами функції є cutoff – задає граничну величину для виділення окремих кластерів шляхом виключення зв'язків дендрограми (рівень зрізу дендрограми); depth – аргумент, що визначає «глибину» ієрархії, для якої підраховуються коефіцієнти несумісності.

Функція `cluster` повертає матрицю ієрархічної структури понять H.

Крок 14. Побудувати дендрограму. Результатом кроку 6 є дендрограма, що побудована по матриці H.

4.4 Модуль інтеграції ієрархічних структур понять

Модуль інтеграції даних реалізований на основі алгоритму інтеграції формалізованих текстів, заснований на методі відображення ієрархічних структур

колекцій текстів (простих онтологій) із застосуванням МРЧ, складається з десяти кроків.

Вхідними даними аналітичного методу інтеграції ІСП є результати, отримані за допомогою аналітичного методу формалізації ЛЕІ (п. 2.1), застосованого для різних колекцій текстів, отриманих у різний час із джерел А и В, але приналежних до одної предметної області: матриці терм-тема $\mathbf{\Phi_A}$ й $\mathbf{\Phi_B}$, тема-документ $\mathbf{\Theta_A}$ і $\mathbf{\Theta_B}$ и ієрархічних зв'язків понять $\mathbf{H_A}$ і $\mathbf{H_B}$. За вхідними даними необхідно знайти подібність двох ієрархічних структур $\mathbf{H_A}$ і $\mathbf{H_B}$ і створити нову ієрархічну структуру H_c, що поєднує і погоджує семантичні представлення вихідних колекцій текстів.

Вихідними даними алгоритму є глобальна матриця семантичної близькості й підсумкова ієрархічна структура понять.

Крок 1. Обчислити семантичну близькість на рівні термів – функція semcos. Вхідними даними є вектор-стовпці $\mathbf{\Phi}^{\kappa}$ матриць $\mathbf{\Phi_A}$ і $\mathbf{\Phi_B}$, отримані із джерел А и В, відповідно. Результатом дії є матриця семантичної близькості термів Stab, тип даних matrix.

Крок 2. Обчислити семантичну близькість на рівні понять – функція seracos. Вхідними даними є вектор-стовпці $\mathbf{\Theta}^{i}$ матриць $\mathbf{\Theta_A}$ і $\mathbf{\Theta_B}$, отримані із джерел А та В, відповідно. Результатом дії є матриця семантичної близькості понять Syeq, тип даних matrix.

Крок 3. Обчислити семантичну близькість на таксономічному рівні функція dep. Вхідними даними є матриці $\mathbf{H}_A$ и $\mathbf{H}_B$ Результатом дії є матриця близькості понять на рівні таксономії $\boldsymbol{S}^{\mathbf{H}}$, тип даних matrix.

Крок 4. Ініціалізувати параметри – функція initswarmmatrix. Початкова популяція часток задається випадково сгенерованим набором значень із обмеженнями $P, \mu, \eta \in [0; 1]$, $P + \mu + \eta = 1$.

Крок 5. Задати цільову функцію – objective. Цільова функція є функцією пристосованості частки.

Крок 6. Виконати МРЧ. Спочатку випадковим образом від 0 до 1 вибирається значення для кожного гнізда. Після того, як обрані початкові рої,

розраховуються відповідні їм значення придатності. Первісна швидкість кожного гнізда частки дорівнює нулю. Реалізація кроку виконується сукупністю функцій, перерахованих у таблиці 4.2.

Таблиця 4.2 – Функції кроку 6, що реалізують метод роя часток

Назва	Опис
creationfcn	Функція створення первинного роя
hybridfcn	Функція, яка продовжує оптимізацію після завершення функції particleswarm.
Initialswarmm	Початкова або часткова популяція часток.
Initialswarmm	Початковий діапазон положень часток.
Maxlterations	Число ітерацій
Objectivelimit	Мінімальне значення цільової функції, критерій зупинки. Скалярний, за замовчуванням -Inf.
Outputfcn	Дескриптор функції або масив гнізд функції. Вихідні функції можуть читати ітеративні дані й зупиняти вирішувач.
Selfadjustmentweight	Зважування кращої позиції кожної частки при настроюванні швидкості.
Socialadjustmentweight	Зважування кращої сусідньої позиції при регулюванні швидкості.
Swarmsize	Кількість часток у рої, ціле число більше 1.
particleswarm	Функція – вирішувач методу роя часток
objective	Дескриптор функції до цільової функції або ім'я цільової функції.
gamultiobj	Функція Паретто
solution	Функція одержання рішення, вектор

Створення роя наступного покоління виконується шляхом оцінки положення й швидкості. Кожне гніздо або позиція являє собою вагу (нормалізоване значення гнізда) щодо міри подоби. Гнізда усередині частки містять значення від 0 до 1, а швидкостям кожного гена задані нульові значення. Використовуючи інформацію, отриману на попередній стадії, положення кожної частки й швидкості кожного кластера обновляються. Кожна частка відслідковує кращу позицію, яку вона досягла ($pbest$). З погляду підходу, позиція вибирається

для кращої позиції, для якої пристосованість цієї частки домінує над іншими пристосуваннями. Найкраще положення серед усіх часток називається глобальним кращим ($gbest$). Далі обчислюється значення цільової функції кожної частки по формулі. На основі отриманих даних проводиться вибір часток, які дають найбільші значення цільової функції. Результатом дії є вектори вагових коефіцієнтів.

Крок 7. Обчислити комплексні заходи семантичної близькості понять – функція $global - sem$ (2.20). Результатом дії є вектор комплексних заходів семантичної близькості понять.

Крок 8. Потім, виконується оцінка комплексних заходів семантичної близькості для понять – функція integfunc (правила 2-4).

Крок 9. Сформувати підсумкову матрицю ІСП – функція $globalmatrix$. Результатом дії є підсумкова матриця ІСП.

Крок 10. Побудувати підсумкову дендрограму – функція $dendrogram$. Результатом дії є дендрограма підсумкової матриці ІСП, побудованої по формалізованих колекціях текстів із джерел А и В.

4.5 Реалізація алгоритму аналізу лінгвістичної експертної інформації

Розроблена проста онтологія інформаційних технологій (ПО ІТ) призначена для застосування в складі систем інтелектуальної обробки текстової інформації, у т.ч. контент-аналіз, інформаційний пошук, побудова баз знань.

Вхідною інформацією є нерозмічені колекції текстів матеріалів міжнародних науково-технічних конференцій «Практичні основи інформаційно-комунікаційних технологій» (AICT) і «Відкриті семантичні технології проектування інтелектуальних систем» (OSTIS) за 2015-2018 рр. Матеріали обрані з рубрик: Інтелектуалізація обробки інформації», «Математичні методи розпізнавання образів», «Штучний інтелект», «Комп'ютерні технології»,

«Програмне забезпечення» і ін. Основними перевагами обраних джерел даних для системи автоматичної побудови простої онтології по колекції текстів є: великий розмір каталогу конференцій; доступність каталогу для завантаження; матеріали конференцій розподілені по рубриках. Виконавцем процесу є експерт предметної області.

Механізмами керування – лінгвістичний процесор, що виконує лінгвістичну розмітку вхідних документів, і розроблений алгоритм інтелектуальної обробки лінгвістичної експертної інформації.

Автоматична побудова простої онтології по колекції текстів здійснюється в шість етапів:

– «Підготовка документів у рамках заданої області наукового знання» – процес полягає у формуванні колекції текстів заданої предметної області. У якості джерела інформації обрані праці наукових конференцій, що перебувають у відкритому доступі в Інтернеті. Матеріали конференцій з'являються щодня, що дозволяє із заданою періодичністю обновляти інформацію про область знання в системі.

Процес виконується експертом вручну. Результатом процесу є одержання експертної інформації:

– «Попередня обробка колекції текстів» – процес полягає в завантаженні документів з колекції текстів і їх лінгвістичній обробці: очищення тексту, графематичний і лексичний аналіз, нормалізація лексем. Процес виконується модулем попередньої обробки колекції текстів. Результатом процесу є одержання лінгвістичної інформації;

– «Витяг термів, тем і понять предметної області» – процес полягає в побудові семантичного ядра текстової колекції, виділенні терміносистеми предметної області – значеннєвих одиниць тексту, що утворюють поняття онтологічної моделі. Процес є автоматичним, виконується модулем формалізації ЛЕІ;

– «Витяг ієрархічних зв'язків між поняттями» – процес полягає в пошуку й виділенні ієрархічних відносин між поняттями простої онтології. Процес є автоматичним, виконується модулем формалізації лінгвістичної експертної інформації;

– «Відображення онтологічної моделі» – виконується висновок отриманої онтологічної моделі у вигляді графа з висновком інформації про кожний елемент онтологічної моделі. На цьому етапі експерт послідовно перевіряє правильність співвіднесення термів предметної області з поняттями в схемі онтології, виділених ієрархічних відносин між темами одного класу. Отримана онтологія зберігається у форматі owl. Процес виконується експертом у системі Protege (етап не автоматизований);

– «Поповнення простої онтології методом інтеграції ІСП» – процес полягає в розширенні семантичного ядра існуючої простої онтології предметної області за рахунок її об'єднання з новою ІСП, отриманої в результаті обробки нової колекції текстів тієї ж предметної області процесами 1-5. Процес об'єднання двох простих онтологій є автоматичним, виконується модулем інтеграції ІСП.

Розроблювальна система дозволяє виконувати напівавтоматичну побудова простої онтології для будь-якої предметної області по колекції текстів. З 3957 різних текстових документів загальним обсягом приблизно 3010000 сторінок тексту без форматування (приблизно 12 Мб файлів формату txt) автоматично було виділено 26860 термів, 254 понять предметної області, 47 ієрархічних зв'язків.

Після завантаження обраних текстів (функція Matlab – extractfiletext) завантажений файл, що містить колекцію документів, відображається в робочій області у вигляді змінної BY2006.mat.

Після попередньої обробки колекції текстів був отриманий одномірний вектор, що полягає з 3957 рядків, кожнен з яких відповідає окремому обробленому документу.

Для побудови тематичної моделі колекції текстів був застосований запропонований у роботі метод, реалізований у вигляді алгоритму.

Вихідна колекція документів була розділена на навчальну й контрольну вибірки. Розбивка виконана за допомогою крос-валідації із застосуванням функції Matlab cvpartition.

Таблиця 4.4 – Фрагмент результатів попередньої обробки документів

№ пп.	Вхідні дані	Вихідні дані	
	Масив textdata	Кількість термів термів	Масив documents
1	"Provides an integrated asset management soultion that will enable USDA to effectively manage it real and personal property assests department-wide."	14	provid integr asset management soultion enabl usda effect manag real person propert assest departmentwid
2	"USDA Corporate finanical information sytems serves as the intitiative for- the OCFO to achieve compliance with the CFO Act of 1990 and- provides the functionality to implement the -US standard General	18	usda corpor finan inform sytem serv intiti ocfo achiev complianc cfo act provid function implement standard gener ledger
3	"Provides an integrated asset management soultion that will enable USDA to effectively manage it real and-personal property assests department-wide."	14	provid integr asset management soultion enabl usda effect manag real person propert assest departmentwid
4	"Aggregation of software applications to assist APHIS in the management of emergencies. Collection of applications includes software for epidemiology, administrative functions and spatial	18	emergen integr collect soult soult funct soft assist applicat **APHIS**
3	**…**		…

Таким чином, була отримана навчальна вибірка (3562 документа, 90%) і контрольна вибірка (395 документів, 10%). По навчальній вибірці була побудована модель «мішка слів». Побудова моделі «мішка слів» для навчальної вибірки виконувалося із застосуванням функції Matlab

(bagofwords(documentstrain)). У підсумку був отриманий «мішок слів» з первісним обсягом словника рівний 3957 слів.

Таблиця 4.4 – Матриця відстаней між поняттями

Теми	Поняття	Індекс теми		
			2	3
1	Intelligence	0	0,04	0,05
2	Analytics	0,04	0	0,12
3	Cyberwar	0,05	0,12	0
4	Technique	0,03	0,06	0,09
5	Software	0,05	0,08	0,14
6	Cybernetic	0,04	0,09	0,13
7	Machine	0,51	0,04	0,05
8	Algorithm	0,05	0,08	0,14
38…				
50	Technique	0,05	0,12	0,37

Отримана матриця відстаней характеризує відстані між окремими темами (поняттями), кожна з яких на першому кроці є окремим кластером.

По отриманій матриці із застосуванням правила об'єднання кластерів середньому зв'язку по мінімальній відстані:

– обчислювалася метрика близькості;

– поєднувалися в новий кластер два найбільше близьких кластера;

– перераховувалися відстані між новим кластером і всіма іншими.

Процес кластеризації тривав до одержання двох останніх кластерів понять. У результаті була отримана матриця подібності, по якій була побудована дендрограма (рис. 4.1).

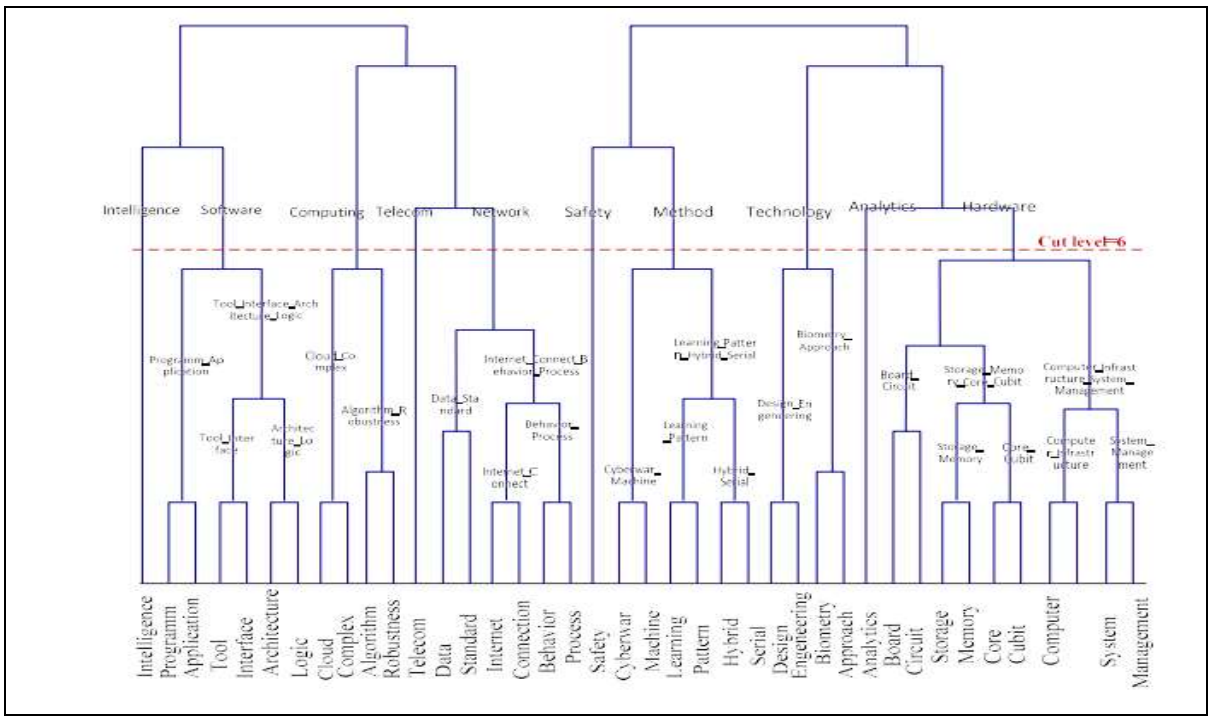

Рисунок 4.1 – Дендрограма ПрО

У процесі кластеризації на шостому кроці було отримано десять кластерів:

– Intelligence / Розвідка;
– Analytics / Аналітика;
– Safety / Безпека;
– Telecom / Телекомунікації;
– Software / ПО (Programmapplicationtoolinterfacearchitecturelogic);
– Computing / Комп'ютеризація (Cloudcomplexalgorithmrobustness);
– Network / Мережа (Datastandard_Internetconnectbehaviorprocess);
– Technology / Технологія (Cyberwarmachinelearningpatternhybridserial);
– Hardware / Апаратура (Boardcircuitcomputer_).

Кожний вузол дендрограми представляє об'єднання двох або більш понять-кластерів, положення понять-вузлів на вертикальній осі визначається відстанню, на якім були об'єднані відповідні поняття-кластери.

5 ОПИС МОЖЛИВОСТІ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Кількість кластерів, при яких досягається однорідне об'єднання в кластери. Кластеризація повинна виявити природні локальні згущення об'єктів. Для вибору кількості кластерів був побудований графік залежності відстані між кластерами від кроку кластеризації.

На графіку на рис. 5.1, перебуває точка «перелому» і номер кроку *m*, на якому відбувся «перелом». У даному випадку точкою перелому є крок під номером 40. Кількість класів рівно *n-m*, де *n* – кількість об'єктів у вибірці. Тому 50-40=10.

Таким чином, підсумкова кількість кластерів дорівнює десяти, що відповідає рівню зрізу дендрограми, що дорівнює 6. Крок кластеризації дорівнює 5 (див. рис.

У результаті онтологічного моделювання розглянутої предметної області засобами редактори Protégé 5.0 була побудована Owl-онтологія (рисунок 5.1), а також граф семантичної мережі.

В ході роботи була створена проста онтологія (таксономія) по предметній області «Інформаційні технології», що має вид орієнтованого графа, вершинами якого є класи, дугами – властивості спрямовані відносини, що представляють зв'язки між ними. У неї входять 50 класів і підкласів. Побудована онтологія є, по суті, каркасом, представленою предметної області «Інформаційні технології». Онтологію можна розширювати, шляхом автоматичної інтеграції таксономічних структур, згідно із запропонованим у роботі методом, додавати нові зв'язки між об'єктами й збільшувати кількість характеристик кожного екземпляра (див. рис. 5.2).

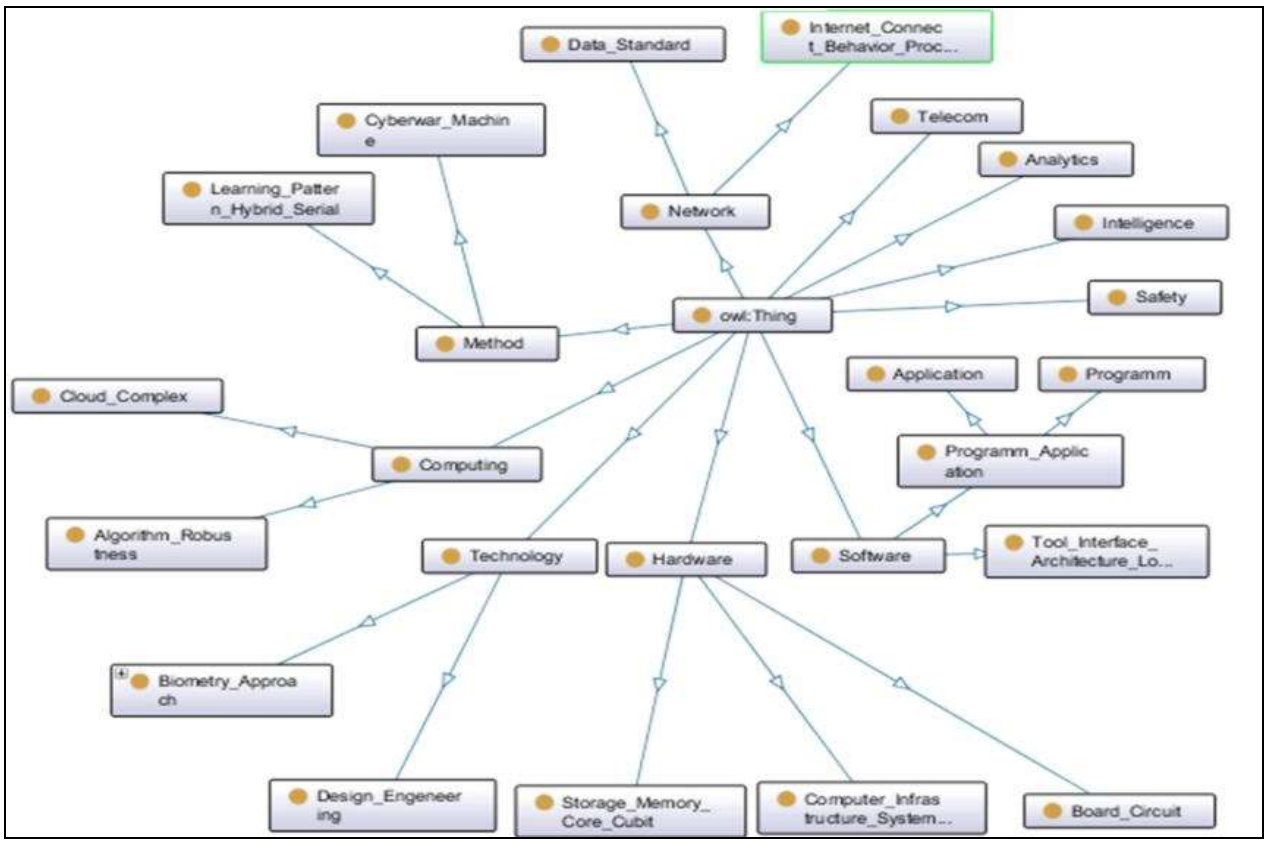

Рисунок 5.1 – Вікно класів простої онтології, побудованої в Protege по отриманій ієрархічній структурі понять

```
<Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-schema#"/>
<Declaration>
   <Class IRI="#Algorithm"/>
</Declaration>
<Declaration>
   <Class IRI="#Application"/>
</Declaration>
<Declaration>
   <Class IRI="#Architecture"/>
</Declaration>
<Declaration>
   <Class IRI="#Automation"/>
</Declaration>
<Declaration>
   <Class IRI="#Board"/>
</Declaration>
<Declaration>
   <Class IRI="#Circuit"/>
</Declaration>
<Declaration>
   <Class IRI="#Cloud"/>
</Declaration>
<Declaration>
   <Class IRI="#Complexity"/>
</Declaration>
<Declaration>
   <Class IRI="#Components"/>
</Declaration>
<Declaration>
   <Class IRI="#Computation"/>
</Declaration>
<Declaration>
   <Class IRI="#Cybersystem"/>
</Declaration>
<Declaration>
   <Class IRI="#Design"/>
</Declaration>
<Declaration>
   <Class IRI="#Distributed"/>
</Declaration>
<Declaration>
   <Class IRI="#Embeded"/>
</Declaration>
<Declaration>
   <Class IRI="#Fault-tolerant"/>
</Declaration>
<Declaration>
   <Class IRI="#Hardware"/>
```

Рисунок 5.2 – Онтологія предметної області (Ovl-формат)

Розроблена онтологія дозволяє оперативно проводити інформаційний пошук у базі знань «Інформаційні технології», визначати взаємозв'язок між онтологічними поняттями, використовуватися для проведення контенту-аналізу.

Аналіз отриманих результатів виконаний шляхом оцінки зовнішніх і внутрішніх критеріїв якості.

Методи оцінки зовнішніх критеріїв пропонованих методів – експертна оцінка. Для проведення експертних опитувань були розроблені аркуші експертних оцінок: оцінка інтерпретованості тем по виділених термінах [12], точність кластеризації, експертна оцінка повноти простої онтології. Групі експертів, що включає фахівців (загальне число опитаних експертів склало 12 людей), що працюють в області інформаційних технологій, були представлені результати реалізації пропонованих методів і алгоритму (виділені термін, теми й поняття, ієрархічне угруповання тем, прості онтології колекцій текстів), які необхідно було оцінити. Завдання експертів – виконати оцінку наступних критеріїв: Інтерпретованість тем, правильність формування поняття щодо змісту теми, число помилок кластеризації, число правильних кластерів. Кожний із критеріїв оцінювався по п'ятибальній шкалі. Для визначення вагових значень критеріїв, експертам також було запропоновано оцінити важливість критерію (метод ранжирування) [16]. За результатами експертної оцінки обчислювалися середнє значення й стандартне відхилення по кожному ваговому критерію, результати аналізувалися й підводили підсумки.

Оцінка внутрішніх критеріїв якості виконана за наступними критеріями: перплексія, когерентність, (погодженість), розмір ядра, контрастність теми, точність знаходження поняття.

Експерименти проводилися на ПК із наступними параметрами: один процесор Intel Core i3-2350M 2,3ГГц, обсяг ОЗУ 8 Гб. Часова складність алгоритму є поліноміальної O(n2).

Характерною точкою можна вважати кількість тем, що дорівнює 50, коли криві метрик когерентності й чистоти тем перетинаються.

При випадковому скороченні словника теми з малою потужністю стають статистично незначущими й перестають виявлятися. Це пояснюється тим, що тематична модель вимушено поєднує основні теми, відмінності між об'єднаними темами стають незначущими й теми зближаються. Аналіз інтерпретованості тем виконаний шляхом експертної оцінки. Кожна тема представлена десятьома найбільш характерними термінами. Тема інтерпретована, якщо за словами, що мають найбільш високу ймовірність теми експерт може визначити, про що ця тема, і дати їй назва.

Приклад інтерпретованої і не інтерпретованої тем. Тема 1 – «Intelligence» (інтелектуальний аналіз даних) включає наступні слова: data, learning, training, model, performance, probability, number, classification, feature, artificial, які можна віднести до області «інтелектуальний аналіз даних». Тема 3 – «Telecom» (Телекомунікаційні системи) характеризується термінами, які не дозволяють однозначно її інтерпретувати: включає багато загальних слів, які можна віднести різним темам, а також трохи виділених помилково, які до теми не ставляться (artificial).

Середня оцінка інтерпретованості тем складає 3,8 ∓ 1,04 балів. Число не інтерпретованих і слабко інтерпретованих тем склало 9 тем; тем, які однозначно інтерпретуються – склало 9 тем.

Таким чином, досліджена правильність вибору оптимальної кількості тем для побудови ієрархічної структури понять ПрО по заданій колекції текстів. На етапі моделювання виділення тем з колекції текстів і побудови їх ієрархічної структури (виконаний багаторазовий прогін навчання моделі, щоб виключити ефект залежності від порядку документів у колекції) були досліджені умови, що впливають на якість виділення понять предметної області. Оптимальні значення метрики досягаються при кількості тем, рівному 50. Ієрархічна кластеризація понять дозволила виділити 10 понять, які стали класами верхнього рівня простої онтології.

ВИСНОВКИ

На підставі проведеного аналізу завдань, пов'язаних з обробкою й аналізом ЛЕІ, отриманої з колекцій текстів заданої предметної області, з метою формалізації й інтеграції, огляду відомих методів була складена таблиця 1.6 проблем аналізу, обробки й формалізації лінгвістичної експертної інформації.

Перші три підзадачі (векторна представлення ЛЕІ, виділення термінів, тем і понять предметної області, побудова ієрархічної структури понять предметної області) етапу формалізації, виконувані з використанням існуючих методів, не дозволяють розв'язати завдання формалізації ЛЕІ повністю автоматичним способом, оскільки вони не погоджені між собою по вхідних і вихідних даних.

Отже, вони не являють собою функціонально закінченої сукупності методів – комплексного методу, орієнтованого на рішення завдання аналізу й обробки ЛЕІ з метою переходу до завдання інтеграції ієрархічних структур понять, розв'язуваних на наступному етапі, що дозволяє в підсумку перейти до побудови простої онтології.

Таким чином, досліджена правильність вибору оптимальної кількості тем для побудови ієрархічної структури понять ПрО по заданій колекції текстів.

На етапі моделювання виділення тем з колекції текстів і побудови їх ієрархічної структури (виконаний багаторазовий прогін навчання моделі, щоб виключити ефект залежності від порядку документів у колекції) були досліджені умови, що впливають на якість виділення понять предметної області. Оптимальні значення метрики досягаються при кількості тем, рівному 50.

Ієрархічна кластеризація понять дозволила виділити 10 понять, які стали класами верхнього рівня простої онтології.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Загорулько Ю.А., Загорулько Г.Б., Боровикова О.І. Технологія створення тематичних інтелектуальних наукових інтернет-ресурсів, що базується на онтології // Програмна інженерія. – 2016. – Т.7, №2. – С. 51–60.

2. Загорулько Г.Б. Розробка онтології для Інтернет-Ресурсу підтримки прийняття рішень у слабоформалізованих областях / Г.Б. Загорулько // Онтологія проектування. – 2016. – Т. 6, №4(22). – С. 485-500. – DOI: 10.18287/2223-9537-2016-6-4-485-500.

3. Зайцева А.С. Моделювання терміносистеми надзвичайних ситуацій / А.С. Зайцева, Ю.В. Сложенікіна // Онтологія проектування – 2018. – Т. 8, №4(30). – С.562-570.

4. Aggarwal Charu C, Zhai Cheng Xiang. Mining text data. Springer Science & Business Media, 2012.

5. Alan Griffiths, H. Claire Luckhurst, and Peter Willett. Using Interdocument Similarity Information in Document Retrieval Systems. Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom. Information Retrieval 1997 p.365-373.

6. Artale A., Franconi E., Guarino N., Pazzi L. Part-Whole Relations in Object-Centered Systems: An Overview // Data and Knowledge Engineering. 1996. V.20. P. 347-383.

7. Whissell John S., Clarke Charles L.A. Improving document clustering using Okapi BM25 feature weighting. Information retrieval. 2011. T. 14, № 5, pp. 513-523.

8. Fouad M. Data mining and fusion techniques for Wsns as a source of the big data / M.M. Fouad, N.E. Oweis, T. Gaber, M. Ahmed, V. Snasel // Procedia Computer Science. – Elsevier, 2015. – Vol.65. – P. 778-786.

9. Hendler A. J. Handbook of Semantic Web Technologies. – Springer, 2019. – 479p.

10. Semantic Web Technologies in Automotive Repair and Diagnostic // URL:

http://www.w3.org/2001/ sw/sweo/public/UseCases/Renault/.

11. RIF basic logic dialect // URL:http://www.w3.org/TR/2013/REC-rif-bld-20130205/.

12. Gruber T. Collective Knowledge Systems: Where the Social Web meets the Semantic Web // Journal of Web Semantics. – 2018. – V. 6, № 1. – P. 4–13.

13. The Description Logic Handbook / F. Baader, D. Calvanese, D. Guinness, D. Nardi, P. Schneider. – Cambridge University Press, 2017. – 574 p.

14. OWL Web Ontology Language Semantics and Abstract Syntax // URL: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/

15. Babenko A., Lempitsky V. Aggregating Deep Convolutional Features for Im- age Retrieval // Proceedings of the IEEE International Conference on Computer Vision. – 2018. P. 1269–1277.

16. Berclaz J., Fleuret F., Fua P. Robust people tracking with global trajectory Computer Vision and Pattern Recognition. – 2016. – P. 744–750.

17. Shubin, I., Snisar, S., Zhyrnov, V., Slavhorodskyi, V.// Practical Application of Formal Representation of Information for Intelligent Radar Systems 2018 International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, PIC S and T 2018 - Proceedings, 2019, pp. 433-436, 8632103

18. Drayer B., Brox T. Object Detection, Tracking, and Motion Segmentation for Object-level Video Segmentation // arxiv.org. 2016. – URL: https://arxiv.org/abs/1608.03066.

19. Hinton G. A practical guide to training restricted Boltzmann machines // Momentum. – 2010. – № 9(1).

20. Kim C., Li F. Multiple Hypothesis Tracking Revisited // Proceedings of the IEEE International Conference on Computer Vision. – 2019.

21. Konev A., Chigorin A., Krivovyaz G., Velizhev A., Konushin A. Traffic signs recognition on images with training on synthetic data // Technical vision in computer systems. – 2019. P. 65-66.

22. Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Kar- pathy A., Khosla A., Bernstein M., Berg A.C., Fei-Fei L. Imagenet large scale visual recognition challenge // IJCV. – 2015

23. Ruta A. A New Approach for In-Vehicle Camera Traffic Sign Detection and Recognition // IAPR Conference on Machine vision Applications (MVA). – 2009. – P. 509-513.

24. Аведьян Є.Д., Галушкин А.І., Селиванов С.А. Порівняльний аналіз структур полносвязных і сверточных нейронних мереж і їх алгоритмів навчання // Інформатизація й зв'язок. – 2017. – № 1.

25. Антощук С.Г. Відстеження об'єктів інтересу при побудові автома- тизированных систем відеоспостереження за людьми // Електротехнічні й комп'ютерні системи. – 2018. – №8(84). – С. 151–156.

26. Chetverikov G., Puzik O., Vechirska I. Multiple-valued structures of intellectual systems //Proceedings of the with Internations Computer Sciences and Information Technologies (CSIT). 2016, 7589907. -pp. 204-207

27. Yang M., Wu Y. A Hybrid Data Association Framework for Robust Online Multi-Object Tracking // arxiv.org 2017. – URL: https://arxiv.org/abs/1703.10764 .

28. Simonyan K., Zisserman A. Very deep convolutional networks for large- scale image recognition // Proceedings of the Neural Information Processing Sys- tems conference, NIPS. – 2015.

29. Szegedy C., Liu W., Jia Y. Going deeper with convolutions // CVPR. – 2015.

30. Talukder K.H., Harada K. Haar Wavelet Based Approach for Image Compres- sion and Quality Assessment of Compressed Image // IAENG International Journal of Applied Mathematics. – 2007. – 36(1).

31. Viola P., Jones M. Robust Real-Time Face Detection // International Journal of Computer Vision. – 2014. – V. 57. – №2. – P. 137–154.

32. A Neural Network for Machine Translation, at Production Scale. URL: https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

33. Добро пожаловать в блог DeepL! URL: https://www.deepl.com/blog/20180215.html

34. Shostak I., Matyushenko I., Romanenkov Yu., Danova M., Kuznetsova Yu. Computer Support for Decision-Making on Defining the Strategy of Green IT Development at the State Level. In book: Green-IT Engineering: Social, Business and Industrial Applications, Vol. 171. Berlin, Heidelberg: Springer International Publishing, 533–559 (2018), https://doi.org/10.1007/978-3-030-00253-4

35. Shostak I., Kapitan R,. Volobuyeva L., and Danova M., Ontological Approach to the Construction of Multi-Agent Systems for the Maintenance Supporting Processes of Production Equipment. In Proc. : IEEE International Scientific and Practical Conference «Problems of Infocommunications. Science and Technology» (PICS&T-2018). Ukraine, Kharkiv, October 9-12, 2018. P. 209 – 214