МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

ДОРОШЕНКО АНАСТАСІЯ ЮРІЇВНА

УДК 004.89:510.635(043.3)

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ФАКТОГРАФІЧНИХ ТЕКСТОВИХ РЕСУРСІВ

Спеціальність 05.13.06 – інформаційні технології

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2019

Дисертацією є рукопис.

Робота виконана на кафедрі інтелектуальних комп'ютерних систем Національного технічного університету «Харківський політехнічний інститут» Міністерства освіти і науки України.

Науковий керівник доктор технічних наук, професор
Шаронова Наталія Валеріївна,
Національний технічний університет
«Харківський політехнічний інститут»,
завідувач кафедри інтелектуальних
комп'ютерних систем.

Офіційні опоненти: доктор технічних наук, професор
Федорович Олег Євгенович,
Національний аерокосмічний університет
імені М.Є. Жуковського «Харківський
авіаційний інститут», завідувач кафедри
інформаційних управляючих систем;

кандидат технічних наук, доцент
Шубін Ігор Юрійович,
Харківський національний університет
радіоелектроніки,
професор кафедри програмної інженерії.

Захист відбудеться «4» квітня 2019 р. о 15 годині 30 хвилин на засіданні спеціалізованої вченої ради Д 64.050.07 в Національному технічному університеті «Харківський політехнічний інститут» за адресою: 61002, м. Харків, вул. Кирпичова, 2.

З дисертацією можна ознайомитись у бібліотеці Національного технічного університету «Харківський політехнічний інститут» за адресою: 61002, м. Харків, вул. Кирпичова, 2.

Автореферат розісланий «4» березня 2019 р.

Вчений секретар
спеціалізованої вченої ради Ю. І. Дорофєєв

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. У зв'язку зі швидким розвитком технологій інформацію розглядають як один із основних ресурсів розвитку суспільства, а інформаційні системи та технології – як знаряддя удосконалення продуктивності праці та ефективності роботи. У будь-яких соціально-економічних та організаційно-виробничих системах опрацювання та переробка інформації є найважливішою функцією, без якої неможлива цілеспрямована діяльність. Обсяги та швидкість інформаційних потоків постійно збільшуються, тому підприємства все частіше звертаються до інтелектуального аналізу як засобу, який дає змогу отримувати корисні відомості з величезної кількості інформації, що зберігається в корпоративних базах даних.

Бурхливий розвиток мережевих інформаційних технологій, в тому числі Інтернету, сприяють значному збільшенню доступних інформаційних ресурсів та обсягів переданої інформації. Найчастіше це різнорідна, слабоструктурована та надлишкова інформація, що має високу динаміку оновлення. Необхідність ефективного використання цього колосального мінливого обсягу інформації обумовлює ***актуальність і значимість*** досліджень у галузі інтелектуальної обробки інформації. Значна частина інформації подається у вигляді текстів природними мовами. У багатьох завданнях, наприклад, при обробці новин або результатів пошуку наукових статей, кількість текстових документів, що вимагають обробки, є дуже значною, тому велику значимість мають методи, які спрощують роботу з такими об'єктами.

Сучасні методи аналізу даних були започатковані та розвинені у працях Ю.П. Адлера, Б.В. Гнеденка, О.Г. Івахненка, Дж. Кіфера, К.Х. Крамера, Б.Ю. Лемешка, Ю.В. Лінника, Г.В. Мартинова, В.В. Налімова, М.С. Нікуліна, О.І. Орлова, Е. Пітмена, Ю.В. Прохорова, Е. Пятецького-Шапіро, Дж. Тьюкі, Г. Хоттелінга, П. Хьюбера, А. Хьютсона, Н.Ф. Хайрової, Ю.П. Шабанова-Кушнаренка, Н.В.Шаронової та багатьох інших дослідників. Останнім часом значного поширення набувають нові технології та методи аналізу даних, зокрема методи інтелектуального аналізу (Data Mining), які використовуються для виявлення прихованих закономірностей у великих масивах даних.

Фактографічна інформація – це інформація, основана на фактах. Автоматизований пошук фактів є одним із найбільш ефективних інструментів ідентифікації інформації для прийняття рішень. Недостатня надійність автоматично вилучених фактів є основною проблемою обробки фактографічної інформації. Наявність різних тлумачень одного і того ж явища, а також неточність або невідповідність інформації, що надходить з різних джерел, підтверджують актуальність досліджень в області лінгвістичного аналізу фактографічної інформації. Аналіз наукових робіт показав, що при існуючому розмаїтті методів інформаційного пошуку проблема залишається недостатньо вирішеною. Одним із перспективних напрямків інформаційного пошуку є фактографічний пошук і розробка фактографічних баз знань. Існуючі сьогодні моделі та алгоритми фактогра-

фічного пошуку у своїй більшості спрямовані на вилучення фактів з добре формалізованої, у тому числі, текстової інформації.

Таким чином, науково-практична задача створення інформаційної технології інтелектуального аналізу фактографічних текстових ресурсів є актуальною та визначає напрям дисертаційного дослідження.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана на кафедрі інтелектуальних комп'ютерних систем НТУ «ХПІ» у межах держбюджетних НДР МОН України: «Розробка математичних моделей та методів розв'язання задач інтелектуальної обробки інформації» (ДР № 0108U003926); «Розробка моделей та методів для інформаційно-пошукових, лексикографічних інтелектуальних систем» (ДР № 0111U002258), у яких здобувач брала участь як виконавець.

Мета і задачі дослідження. Метою дисертаційної роботи є забезпечення несуперечливості та актуальності результатів інтелектуального аналізу фактографічних ресурсів за рахунок зменшення часу пошуку фактографічної інформації та нормалізації результатів пошуку на основі розробки алгебро-логічних моделей екстракції фактів та відповідної інформаційної технології.

Відповідно до зазначеної мети поставлено наступні **задачі**:

1. Проаналізувати існуючі інформаційні технології, моделі та методи обробки фактографічних даних у мережевих потоках слабоструктурованої текстової інформації, сформулювати основні вимоги до розробки інформаційної технології інтелектуального аналізу фактографічних ресурсів.

2. Обґрунтувати вибір математичного інструментарію для моделювання процесу екстракції фактографічних даних на основі інтелектуального аналізу мережевих текстових фактографічних ресурсів.

3. Розробити моделі тематичного пошуку та екстракції фактографічної інформації на основі інтелектуальної процедури оцінки релевантності текстової інформації.

4. Сформувати підхід до видобування фактографічних даних з текстових джерел на основі використання онтологічних специфікацій та розширення онтологічної моделі предметної області.

5. Розробити інформаційну технологію інтелектуального аналізу фактографічної інформації.

6. Провести апробацію розроблених моделей, підходів та інформаційної технології та впровадити результати дослідження в існуючі нформаційні системи.

Об'єктом дослідження є процес інтелектуального аналізу фактографічної інформації у текстових даних.

Предметом дослідження є моделі та інформаційна технологія пошуку, збору та ідентифікації фактографічної інформації в текстових мережевих ресурсах.

Методи досліджень базуються на використанні теорії інтелекту та методу компараторної ідентифікації, які комплексно застосовуються при створенні

інформаційно-логічних моделей інтелектуальної обробки фактографічної текстової інформації. Використовується математичний апарат алгебри логіки, теорії категорій, теорії відношень, методи класифікації та кластеризації, методи інтеграції та пошуку інформації на основі онтологій. Для розробки концептуальної моделі інтелектуального аналізу фактографічних текстових ресурсів використано метод компараторної ідентифікації та апарат алгебри скінченних предикатів.

Наукова новизна отриманих результатів:

– *вперше* запропоновано розв'язання задачі інтелектуального аналізу мережевих текстових фактографічних ресурсів на основі використання інструментів алгебри скінченних предикатів і методів теорії інтелекту, що дозволило скоротити час та збільшити точність тематичного інформаційного пошуку за смисловими характеристиками фактів;

– *отримав подальший розвиток* метод екстракції фактографічних даних, заснований на методі компараторної ідентифікації, що дозволяє вирішити задачу моделювання інформаційного пошуку на основі семантичного оцінювання триплетів фактів;

– *удосконалено* метод розширення онтологій для опису процесів інтеграції фактографічної інформації, що дозволяє збільшити виразні спроможності засобів опису предметних областей для проведення їх автоматизованого аналізу на основі використання предикатної інтерпретації категорій;

– *удосконалено* інформаційну технологію інтелектуального аналізу текстової інформації, яка дозволяє автоматизувати процес обробки слабоструктурованих фактографічних ресурсів та вдосконалити процес екстракції знань за рахунок визначення ознак та атрибутів предметних областей (ПО).

Практичне значення одержаних результатів полягає у розробці інформаційної технології інтелектуального аналізу текстової інформації, яка дозволяє автоматизувати процес обробки фактографічних ресурсів та вдосконалити процес екстракції знань за рахунок визначення ознак та атрибутів ПО, побудови відповідних онтологій для ПО «радіаційна безпека» та «обробка текстових патентно-кон'юнктурних даних».

Отримані в роботі результати знайшли практичне застосування у системах збору, ідентифікації та екстракції фактографічної інформації та знань у Наукових бібліотеках: Харківського національного університету радіоелектроніки (ХНУРЕ) для удосконалення засобів обробки фактографічної текстової інформації (довідка від 08.10.2018 р.); Національного юридичного університету імені Я. Мудрого для підвищення якості обробки інформації в автоматизованій інформаційно-пошуковій бібліотечній системі за рахунок зменшення часу пошуку фактографічних ресурсів, поданих природною мовою (довідка від 26.09.2018 р.). Результати роботи використані у ТОВ Харківський Технічний Центр Оцінки «Експертус» для збору маркетингової інформації з веб-ресурсів (довідка від 10.10.2018 р.). Теоретичні результати дисертації використовуються в навчальному процесі на кафедрі інтелектуальних комп'ютерних систем НТУ

«ХПІ» при викладанні спеціальних дисциплін «Інформаційно-ресурсне забезпечення лінгвістичної діяльності», «Сучасні інтелектуальні системи аналізу та обробки інформації» у розділі «Екстракція знань та пошук фактографічної інформації» для студентів спеціальності «Прикладна лінгвістика» та при виконанні курсових та дипломних робіт (довідка від 06.11.2018 р.).

Особистий внесок здобувача. Усі основні результати дисертаційної роботи, що виносяться на захист, отримані здобувачем особисто. Серед них: огляд сучасних методів інтелектуального аналізу фактографічної інформації та концептуальна модель інтелектуального аналізу фактів; метод використання онтологій для семантичного пошуку документів; метод розширення онтологій для інтеграції фактографічних ресурсів; засоби обробки патентно-кон'юнктурної інформації (ПКІ) на основі онтологій; метод обробки фактографічної інформації для текстової ПКІ; розробка програмних компонентів інформаційної системи екстракції фактографічних даних з веб-ресурсів.

Апробація результатів дисертації. Результати дисертаційної роботи доповідались та обговорювались на міжнародних та всеукраїнських наукових конференціях: «Computer Science and Information Technologies» (Львів, 2008); «Інтелектуальні системи та прикладна лінгвістика» (Харків, 2012); «Інформаційні проблеми теорії акустичних, радіоелектронних і телекомунікаційних систем IPST» (Алушта, 2013); «Експертні оцінки елементів навчального процесу» (Харків, 2013); «Актуальні проблеми прикладної лінгвістики очима наукової молоді» (Харків, 2013); «Computational Linguistics and Intelligent Systems (COLINS)» (Львів, 2018); «Інформаційні системи і технології» (Коблево, 2018).

Публікації. Основні результати дисертації опубліковані у 18 наукових працях, у тому числі 5 статей у наукових фахових виданнях України (4 – у виданнях, що входять до міжнародних наукометричних баз), 2 патенти, 11 – у матеріалах конференцій (1 стаття проіндексована у Scopus).

Структура та обсяг дисертації. Дисертаційна робота складається з анотації двома мовами, вступу, 4 розділів, висновків, списку використаних джерел та додатків. Повний обсяг дисертації складає 178 сторінок, з них 30 рисунків по тексту, 9 таблиць по тексту, список з 175 найменувань використаних джерел на 20 сторінках, додатки на 9 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертації, зазначено зв'язок роботи з науковими темами, сформульовано мету і задачі дослідження, визначено об'єкт, предмет і методи дослідження, показано наукову новизну та практичне значення отриманих результатів, наведено інформацію про практичне використання, апробацію результатів та їх висвітлення у публікаціях.

У **першому розділі** проведено аналіз та класифікацію задач пошуку текстової інформації, здійснено огляд проблем, переваг та недоліків існуючих методів інформаційного пошуку. Проведено аналіз досліджень в області інтелек-

туального тематичного аналізу текстів, розглянуто задачі класифікації та кластеризації, визначено необхідність досліджень засобів математичного моделювання інформаційної відстані, наведено постановку задач в області інформаційних технологій обробки фактографічних даних у мережах.

Крім обміну інформацією з іншими користувачами, для користувачів Інтернету важливим є пошук необхідної інформації за запитом, що відповідає суб'єктивним критеріям, у зв'язку з чим виникає завдання інформаційного пошуку (ІП) (англ. Information Retrieval) як процесу пошуку неструктурованої інформації. Інформація, яка характеризує певний конкретний факт, фактографічну подію або їх сукупність, називається фактографічною. Для того, щоб спростити пошук інформації і зробити його релевантним, застосовується певна кількість методів ІП, але проведений огляд проблем, переваг та недоліків існуючих методів ІП показав, що при їх великої кількості проблема автоматизованого пошуку фактографічної інформації є недостатньо вирішеною.

Розглянуто онтологічний підхід, який дозволяє на основі семантичного опису ресурсів знань інформаційного простору специфікувати та побудувати пошуковий образ запиту. Застосування онтології дозволяє підвищити пертинентність ІП, тобто відповідність отриманих ресурсів певній інформаційній потребі. Застосування методів інтеграції та пошуку інформації на основі онтології використовується для забезпечення підтримки управлінських рішень і є засобом представлення семантики.

Таким чином, у першому розділі проаналізовано існуючі інформаційні технології, моделі та методи обробки фактографічних даних у мережевих потоках слабо структурованої текстової інформації, сформульовано основні вимоги до розробки інформаційної технології інтелектуального аналізу фактографічних ресурсів, зроблено постановку задач дослідження.

У другому розділі проаналізовано особливості інтелектуального аналізу фактографічної текстової інформації. Обґрунтовано вибір математичного інструментарію для моделювання процесу обробки фактів, описано базові засади та принципи моделювання лінгвістичної обробки фактографічних ресурсів, а також базові моделі інтелектуальної обробки фактографічної інформації.

Реалізація ефективного пошуку фактографічної інформації вимагає вивчення структури предметної області, знаходження її специфічних семантичних ознак, дослідження процесу пошуку релевантних джерел фахівцем в галузі. Формальне представлення фактографічної інформації можливе лише на основі моделювання дії фахівця при аналізі повнотекстової інформації та ідентифікації її змісту. Для цього використовується метод компараторної ідентифікації лінгвістичних об'єктів, який є ефективним засобом опису інтелектуальної діяльності людини. Теорія компараторної ідентифікації дозволяє з'ясовувати внутрішню структуру інформаційних сигналів, вигляд функції перетворення змісту інформації та вигляд предикату, який описує вибір дії фахівцем. Ще більш абстрактним і потужним інструментом, який використовується для потреб інформатизації, у тому числі для машинного подання й обробки знань, є теорія категорій.

Надано характеристику поняття класичної категорії як сукупності однотипних математичних об'єктів (множин, просторів, груп і т.д.) та їх відображень один на одного (морфізмів). Клас об'єктів категорії K позначається $Ob\,K$, а клас морфізмів – $Mor\,K$. Безоб'єктна класична категорія є одним із видів алгебри та задається множиною M, елементи якої називаються морфізмами, та єдиною частковою бінарною операцією $f°g$ множення морфізмів, яка відображає декартовий добуток $M \times M$ у M. Добуток морфізмів є асоціативним $(f°g)°h = f°(g°h)$ для будь-яких $f, g, h \in M$, для яких існують добутки $(f°g)°h$ та $f°(g°h)$. Множина M морфізмів з одиничними морфізмами та з діючою на ній операцією множення, яка має вище наведені властивості, називається безоб'єктною категорією K.

В об'єктній категорії додатково до морфізмів введено поняття об'єктів. Об'єкти позначаються буквами $A, B, C, \ldots$ Якщо $A \in Ob\,K$, то A є об'єктом категорії K або K -об'єктом. Говорять, що f є морфізмом із об'єкта A в об'єкт B та пишуть $f: A \to B$ або $A \xrightarrow{f} B$. Над множиною $Mor\,K$ визначено часткову двомісну операцію множення морфізмів. Добуток $f°g$ морфізмів $f: A \to B$ та $g: C \to D$ є визначеним тільки в тому випадку, якщо $B = C$, тобто кінець морфізму f співпадає з початком морфізму g. У цьому випадку добуток $f°g$ є морфізмом з об'єкта A в об'єкт D. У цьому випадку для об'єктів $A, B, C \in K$ визначено відображення $HK(A,B) \times HK(B,C) \to HK(A,C)$. Множення морфізмів є асоціативним $(f°g)°h = f°(g°h)$, коли $f: A \to B, g: B \to C, h: C \to D$.

Класична категорія допускає різні інтерпретації, у тому числі проективну та предикатну інтерпретації. Предикатна інтерпретація категорії задається на деякому універсумі U. У ролі об'єктів $A, B, C, \ldots$ використовуються довільні підмножини універсуму U. У ролі морфізму $f: A \to B$ використовується довільний лінійний логічний оператор $F: f(P) = Q$, який перетворює предикат P у предикат Q та записується у вигляді: $\exists x \in A(K_f(x,y)P(x)) = Q(y)$. Предикат $P(x)$, заданий на множині A, розглядається як екземпляр об'єкта A, предикат $Q(y)$, заданий на множині B, – як екземпляр об'єкта B. Морфізм $f: A \to B$ перетворює екземпляри об'єкта A в екземпляри об'єкта B. Предикат $K_f(x,y)$ є ядром лінійного логічного оператора F_f, який повністю характеризує відповідне перетворення. Предикат $K_f(x,y)$ заданий на декартовому добутку $A \times B$ множин A та B. Морфізм f повністю заданий предикатом $K_f(x,y)$. У ролі множини $Mor(A,B)$ обрано систему різноманітних операцій. У категорії $Pred$ кожному морфізму $f \in Pred$ взаємно однозначно відповідає ядро $K_f(x,y)$. Кожний морфізм $f: \mathrm{A} \to B$ категорії $Pred$ задається предикатом $K_f(x,y)$, який заданий на $A \times B$. Множину $Mor\,Pred$ отримують при об'єднанні множин $Mor\,Pred(A,B)$, де (A,B) – різноманітні пари із множин $A, B \subseteq U$.

Розбір речень у текстових фактографічних ресурсах здійснюється на основі введення предметних змінних. Наприклад, для речення «*Висока частота*

поширеності захворювань потребує удосконалення знань у цій галузі» предметними змінними є самі слова цього речення. Прикладом ядра морфізму, заданого на декартовому добутку $A \times B$ множин $A = \{\text{висока}, \text{середня}, \text{низька}\}$ та $B = \{\text{частота}, \text{амплітуда}, \text{швидкість}\}$, є предикат

$$K_f(x, y) = x^{\text{в}}(y^{\text{ч}} \vee y^{\text{ш}}) \vee x^{\text{с}}(y^{\text{ч}} \vee y^{\text{а}}) \vee x^{\text{н}}(y^{\text{ч}} \vee y^{\text{ш}}).$$

Двочастковий граф предикату $K_f(x, y)$ представлений на рис. 1. Лінійний логічний оператор із цим ядром записується таким чином

$$Q(y) = \exists x \in \{x^{\text{в}}, x^{\text{с}}, x^{\text{н}}\}\big(x^{\text{в}}(y^{\text{ч}} \vee y^{\text{ш}}) \vee x^{\text{с}}(y^{\text{ч}} \vee y^{\text{а}}) \vee x^{\text{н}}(y^{\text{ч}} \vee y^{\text{ш}})\big)P(x).$$

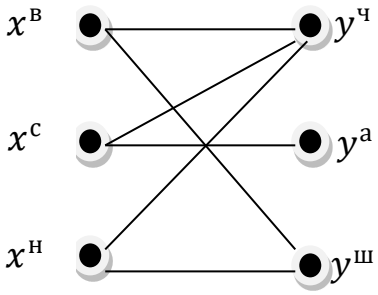

Рисунок 1 – Двочастковий граф предиката $\text{K}_\text{f}(\text{x}, \text{y})$

Якщо замість предиката $P(x)$ підставити у формулу його конкретне значення, наприклад $P(x) = x^{\text{в}}$, то в результаті отримаємо таке значення предиката $Q(y)$

$$Q(y) = \exists x \in \{x^{\text{в}}, x^{\text{с}}, x^{\text{н}}\}\big(x^{\text{в}}(y^{\text{ч}} \vee y^{\text{ш}}) \vee x^{\text{с}}(y^{\text{ч}} \vee y^{\text{а}}) \vee x^{\text{н}}(y^{\text{ч}} \vee y^{\text{ш}})\big)x^{\text{в}} = y^{\text{ч}} \vee y^{\text{ш}}.$$

Цей результат можна одержати також і графічно, якщо елементам множини $P = \{x^{\text{в}}\}$ за допомогою ребер двочасткового графа предиката $K_f(x, y)$ поставити у відповідність пов’язані з ними елементи множини $Q = \{y^{\text{ч}}, y^{\text{ш}}\}$. Таким чином, морфізм $Q(y)$ перетворює множину $P = \{x^{\text{в}}\}$ у множину $Q = \{y^{\text{ч}}, y^{\text{ш}}\}$. Даний приклад ілюструє можливість використання морфізмів предикатної категорії для зберігання знань про те, які словосполучення природної мови можуть бути утворені на множинах слів $A = \{\text{висока}, \text{середня}, \text{низька}\}$ та $B = \{\text{частота}, \text{амплітуда}, \text{швидкість}\}$, а також для виконання запитів типу «*Які словосполучення утворять слова, якщо на першому місці знаходиться слово «*висока*»?*». У цьому прикладі відповіддю на запит є словосполучення: «висока частота», «висока швидкість». Ядро лінійного логічного оператора розглядається як знання або правила одержання знань, а сам лінійний логічний оператор – як механізм виконання запиту для отримання нових знань.

Теорія категорій займається вивченням зв’язку властивостей об’єкта з його внутрішньою структурою і виражає структуру математичного об’єкта за його властивостями. Історично засоби теорії категорій використовувалися в області

математичного опису баз даних. Теорія категорій дає можливість ясно й наочно описувати процеси формування та обробки знань. Із цією метою в роботі використовується алгебра предикатів, засобами якої побудовано предикатну інтерпретацію категорії. Алгебра предикатів претендує на роль універсального математичного засобу для формального опису інформаційних процесів. Наявність алгебри скінченних предикатів відкриває можливість переходу від алгоритмічного опису інформаційних процесів до опису їх у вигляді рівнянь, які задають відношення між змінними. Усі змінні в рівнянні рівноправні, при цьому рівняння мають перевагу перед алгоритмами, оскільки дозволяють розрахувати реакцію системи навіть при неповній визначеності вхідних сигналів. Таким чином, у роботі алгебра скінченних предикатів розглядається як інструмент дослідження. Запропоновано використання предикатних категорій для формалізації фактографічної інформації. Розглянуто та побудовано відповідні реляційні моделі семантичних зв'язків елементів фактографічної інформації за допомогою алгебри предикатів.

У третьому розділі наведено еталонну модель аналізу фактографічної інформації, розроблено моделі інформаційного пошуку фактів. Проведено аналіз та класифікацію онтологій з метою використання онтологічного підходу до опису процесів інтеграції фактографічної інформації.

Розглянуто особливості інтелектуального аналізу фактографічної текстової інформації. Наведено класифікацію фактів та етапи виділення фактів зі слабо структурованої текстової інформації. Запропоновано для опису фактів використання двох типів триплетів: «Суб'єкт→Предикат→Об'єкт» та «Предмет→Атрибут→Значення». Це дозволяє вилучати поняття зі слабоструктурованих текстових ресурсів і описувати відношення між ними у структурованому вигляді. Запропоновано еталонну модель аналізу фактографічної інформації (рис. 2). Здійснено моделювання інформаційного пошуку фактографічної інформації на основі математичної моделі компаратора.

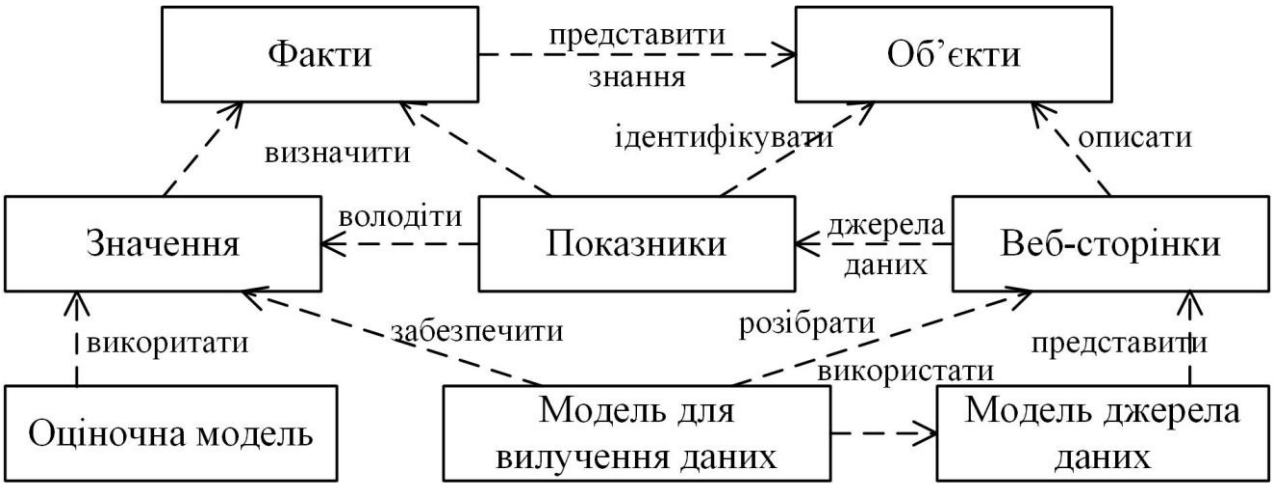

Рисунок 2 – Еталонна модель аналізу фактографічної інформації

Пошук факту – це пошук у семантичній мережі тексту такої підмережі, яка є ізоморфною до одного з шаблонів. Якщо підмережа знайдена, факт вважається встановленим, після чого здійснюється вилучення сутностей та їх маркування ролями, які задані у відповідних вузлах лінгвістичного опису. Інформаційна система, яка вирішує задачі пошуку та аналізу фактографічної інформації, потребує базу знань, яка відображає основні співвідношення понять у певній предметній галузі. Відомості про ці відношення можуть бути використані при побудові тезауруса та онтології предметної галузі. За допомогою предикатних категорій множина правил виводу зберігається у вигляді ядр лінійних операторів, а сам механізм формування знань – у вигляді лінійних операторів, представлених за допомогою формул алгебри предикатів.

Вирішення задач збору фактографічної інформації базується на моделях інформаційного фактографічного пошуку та екстракції даних. Моделі інформаційного пошуку фактографічної інформації задаються на базі компаратора. Розглянемо ці моделі більш детально. Нехай E – множина структурних елементів веб-сторінки, W – множина слів. Тоді $R_{SEARCH} \subseteq E \times W$ – бінарне відношення «використовується для пошуку». Нехай $E_q \subseteq E$ – множина елементів веб-сторінки, які обрані для оцінки та $W_p \subseteq W$ – множина слів, які відповідають темі пошуку. Бінарне відношення $R_{SEARCH} = \{(e_{qi}, w_{pi})\}$ задає пари «елемент–слово», у яких слова належать множині слів, що відповідають темі, та елементи належать множині елементів, обраних для розгляду.

Нехай $w_{pi} \in W_p$ – множина слів, видобутих із веб-сторінки. Тоді предикат, який оцінює бінарні пари «елемент–слово», має вигляд

$$P_w(e_{qi}, w_{pj}) = \begin{cases} 1, \text{ якщо } (e_{qi}, w_{pj}) \in R_{SEARCH}, \\ 0, \text{ якщо } (e_{qi}, w_{pj}) \notin R_{SEARCH}. \end{cases}$$

Предикат, який визначає наявність контрольних слів в певному елементі, має вигляд

$$P_e(e_{qi}) = P_w(e_{qi}, w_{p1}) \vee P_w(e_{qi}, w_{p2}) \vee \ldots \vee P_w(e_{qi}, w_{pn}).$$

Оцінка веб-сторінки об’єднує оцінки за кожним елементом та визначається предикатом

$$P_q = P_e(e_{q1}) \vee P_e(e_{q2}) \vee \ldots \vee P_e(e_{qs}).$$

Для подання моделі шаблону задається бінарне відношення «елементи та відповідні слова, які було видобуто зі сторінки-джерела» $R_{PAGE} \subseteq E \times W$, $R_{PAGE} = \{(e_1, w_1), \ldots, (e_s, w_j)\}$. Функція перетворення комбінацій слів на шаблонні значення із множини «еталонів» $C = \{c_1, \ldots, c_m\}$ задається таким чином

$$\forall(e_i, w_j) \in R_{PAGE}: F(e_i) = \begin{cases} c_1, \text{ якщо } (w_{i1} \wedge w_{i2} \wedge \dots) \vee (w_{j1} \wedge w_{j2} \wedge \dots) \vee \dots \\ \dots \\ c_m, \text{ якщо } (w_{im} \wedge w_{im} \wedge \dots) \vee (w_{jm} \wedge w_{jm} \wedge \dots) \vee \dots \end{cases}$$

Множина елементів веб-сторінки, яка містить певний еталон із ни $C = \{c_1, \dots, c_m\}$, задається як: $E_p = \{e_j \in E | c = F(e_j), c \in C\}$. Нехай $R_{PATTERN} \subseteq E \times C$ – бінарне відношення «елементи містять еталони», при цьому му $R_{PATTERN} = \{(e_i, c_j) | e_i \in E_p, c_j \in C\}$.

Предикат шаблону має вигляд

$$P_{PATTERN} = \begin{cases} 1, \text{якщо} \left(\exists e_1 \exists e_2 \exists e_3 \left(E(e_1, e_F) \wedge E(e_2, e_I) \wedge E(e_3, e_O)\right)\right) = 1, \\ 0, \text{в іншому випадку,} \end{cases}$$

де $E(e_1, e_F) = \begin{cases} 1, e \in E_F, \\ 0, e \notin E_{F,} \end{cases}$ $E(e_2, e_I) = \begin{cases} 1, e \in E_I, \\ 0, e \notin E_I \end{cases}$ та $E(e_3, e_O) = \begin{cases} 1, e \in E_O, \\ 0, e \notin E_{O.} \end{cases}$

Якщо кожен об’єкт описується властивостями або ознаками, то він представляється як точка в n-мірному просторі, і схожість з іншими об’єктами визначається як відповідна відстань. При класифікації об’єктів використовуються різні міри відстані, які детально розглянуто в розділі.

Розглянуто два типи онтологій: семантична та прагматична. Описано використання онтологій для опису процесів інтеграції фактографічної інформації. Основу предметної області складають онтології, що використовуються для опису знань з певної галузі. Узгодження онтологій є вирішенням проблеми семантичної неоднорідності, що є важливим для наступних завдань: розвиток онтологій; інтеграція схем; інтеграція каталогів; інтеграція даних; відповідь на запити тощо. Над онтологіями здійснюються наступні операції: 1) узгодження онтологій – це процес пошуку відношень або відповідностей між сутностями різних онтологій; 2) співвідношення онтологій – це множина відповідностей між двома та(або) більше онтологіями; 3) перетворення онтологій – відповідність сутностей двох онтологій; 4) об’єднання онтологій – це процес створення нової онтології на основі двох вихідних онтологій, що, можливо, перекриваються; 4) інтеграція онтологій – це процес включення однієї онтології до складу іншої таким чином, щоб інтегрована онтологія містила інформацію з обох онтологій; 5) переклад онтології – це процес трансформації онтології з однієї онтологічної мови до іншої; 6) переклад даних – це процес трансформації екземплярів з однієї онтології у відповідні дані або екземпляри іншої онтології.

Процес узгодження онтологій розглядається як функція f двох онтологій O та O', що потребують узгодження, вхідного співвідношення A множини параметрів p, множини передбачень та ресурсів r, що повертає співвідношення A' між цими онтологіями

$$A' = f(O, O', A, p, r).$$

Для двох онтологій O та O' співвідношення – це множина відповідностей між парами сутностей, що належать до онтологій O та O' , відповідно. Для двох онтологій O та O' відповідність M між O та O' – це кортеж $\langle id, e, e', R, n \rangle$, де id– унікальний ідентифікатор відповідності; e та e' — сутності онтологій O та O' відповідно; R – відношення; n — рівень упевненості (зазвичай в інтервалі [0,1]).

Сформовано підхід до видобування фактографічних даних з текстових джерел на основі використання онтологічних специфікацій (рис. 3). Описано використання онтологій для опису процесів інтеграції фактографічної інформації. Зазначено, що онтологія має ґрунтуватися на перевірених джерелах знань, а також передбачати повторне використання вже існуючих онтологій для того, щоб уникати дублювання інформації. Розглянуто приклад формування шаблонів для вилучення фактографічної інформації з текстів англійською мовою. Позначено $\sigma = \{\sigma_0, \sigma_{num}, \sigma_{gend}\}$ – група іменника, де $\sigma_0 -$ начальна форма; $\sigma_{num}, \sigma_{gend}$ – характеристики (число, рід). Групу дієслова (інфінітив, активний стан, пасивний стан, герундій, з прийменником) позначено через $\mu = \{\mu_0, \mu_{act}, \mu_{pas}, \mu_{grd}, \mu_{prep}\}$. Виділено наступні типові лінгвістичні шаблони, які зустрічаються у тексті, та можливі способи онтологічного їх подання, наприклад: (σ, μ_{act}); $(\sigma_0, \mu_{act}, \sigma)$; (σ, μ_{pas}); (σ, μ_i, μ_0); (σ, μ_0, μ_i); (μ_0, μ_i, σ); (μ_0, σ); (μ_{grd}, σ); $(\sigma_i, \mu_{prep}, \sigma_{j0})$; $(\mu_{iact}, \mu_{jprep}, \sigma)$; $(\mu_{i\,pas}, \mu_{j\,prep}, \sigma)$; $(\mu_{i0}, \mu_{jprep}, \sigma)$; (σ_0, σ).

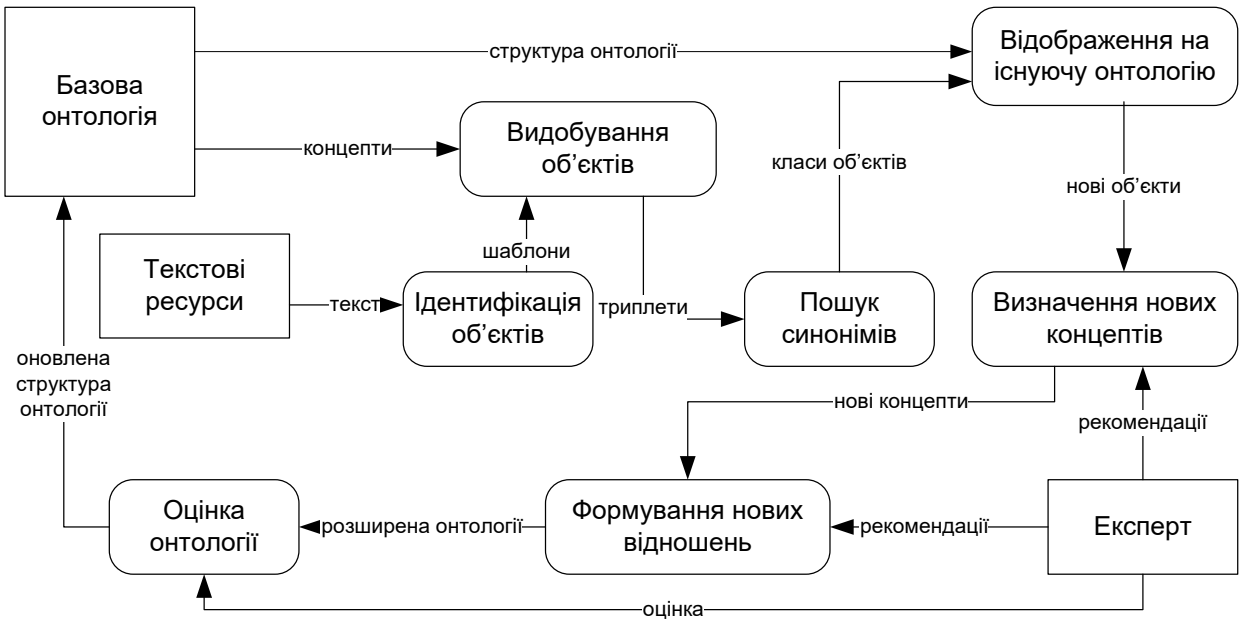

Рисунок 3 – Діаграма потоків даних процесу обробки фактографічної інформації на основі онтологічних специфікацій

Позначено $\Omega = (\psi, \varepsilon, \zeta, \varphi)$ – це базова онтологія, де ψ – множина концептів, ε – множина відношень, ζ – множина допустимих атрибутів (які задаються

ім'ям та типом), та φ – правила виведення, $\psi_s \in \psi$, де $\psi_0 \in \psi$ – суб'єкт та об'єкт відношення, $\tilde{\varepsilon}$ – відношення успадкування. Тоді типові лінгвістичні шаблони матимуть відповідні онтологічні трактовки: 1) $\exists \varepsilon, \psi_s \in \Omega \mid \exists \varepsilon(\psi_s)$; 2) $\exists \varepsilon, \zeta_i \in \Omega \mid \exists \varepsilon(\psi_s) \wedge \zeta_i \epsilon \psi_s$; 3) $\exists \psi_i, \zeta_i \in \Omega \mid \zeta_i \epsilon\, \psi_i$; 4) $\exists \varepsilon, \psi_s, \psi_0 \in \Omega \mid \exists \varepsilon(\psi_s, \psi_0)$; 5) $\exists \varepsilon, \psi_0 \in \Omega \mid \exists \varepsilon(\psi_0)$; 6) $\exists \varepsilon, \psi_i, \psi_j \in \Omega \mid \exists \varepsilon(\psi_i, \psi_j)$; 7) $\exists \varepsilon, \zeta_i \in \Omega \mid \exists \psi_s \exists \varepsilon(\psi_s) \wedge \zeta_i \epsilon \psi_s$; 8) $\exists \varepsilon\ ,\ \zeta_i \in \Omega \mid \exists \psi_0 : \exists \varepsilon\ (\psi_0) \wedge \zeta_i \in \psi_0$; 9) $\exists \varepsilon_i,\ \varepsilon_j,\ \psi_s \in \Omega \mid \exists \varepsilon_i\ (\psi_s) \wedge \exists\ \varepsilon_j\ (\psi_s)$;; 10) $\exists\ \varepsilon_i,\ \varepsilon_j,\ \psi_0 \in \Omega \mid \exists \varepsilon_i\ (\psi_0)\ \wedge\ \exists\ \varepsilon_j(\psi_0)$; 11) $\exists\, \tilde{\varepsilon},\ \psi_i,\ \psi_j \in \Omega \mid \exists \tilde{\varepsilon}\ (\psi_i, \psi_j)$; 12) $\exists\ \psi_i, \zeta_j \in \Omega \mid \zeta_j \in \psi_i$.

Особлива увага приділяється етапу перевірки онтології шляхом побудови семантичних дескрипторів документів та аналізу протиріч, оскільки він є критичним для всієї процедури побудови онтології та є основною відмінністю запропонованого підходу в порівнянні з відомими методами, при цьому він не є незалежним етапом, а постійним процесом автоматичної корекції та верифікації, що запускається після кожного з етапів. Вводиться метрика коректності для синтаксичного $p_{syn}(w_i, w_j, \Omega)$ та семантичного зв'язку $p_{sem}(\alpha_i, \alpha_j, \Omega)$, яка визначає, наскільки коректним є побудований зв'язок між концептами α_i та α_j й також, відповідно, групами слів, які представляють їх в тексті $\{w_i\} \to \alpha_i$ і $\{w_j\} \to \alpha_j$.

Таким чином, інформаційна технологія інтелектуального аналізу фактографічної інформації удосконалює та доповнює існуючий підхід обробки текстових даних і не суперечить існуючій практиці, що свідчить про її практичну цінність та ефективність використання. Схему розробленої інформаційної технології представлено на рис. 4.

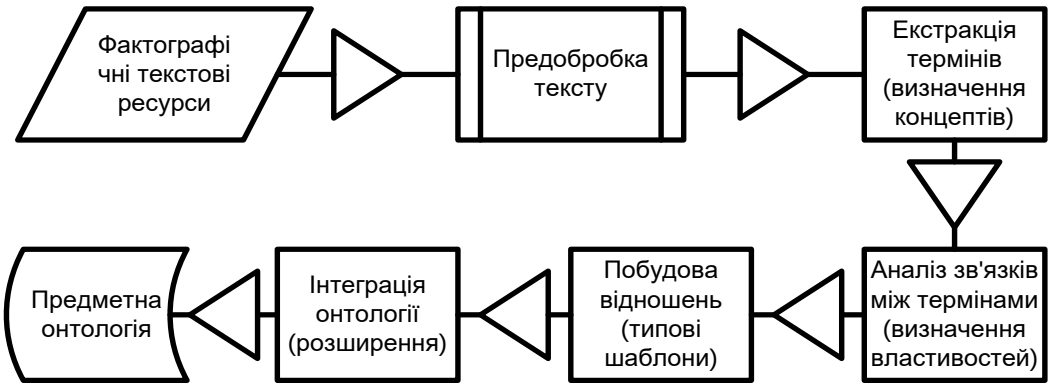

Рисунок 4 – Схема інформаційної технології інтелектуального аналізу фактографічної інформації

Четвертий розділ присвячений практичній реалізації результатів дослідження. Проведено аналіз проблеми та особливості практичної реалізації вирішення задачі екстракції фактографічних даних, розглянуто підходи та інформаційні технології вирішення задач парсингу на базі існуючих інформаційних систем. У дисертаційній роботі запропоновано моделі пошуку, екстракції та об-

робки фактографічних даних на основі комплексу логіко-лінгвістичних моделей. Обробка текстових даних вимагає багато часу, а зберігання її результатів потребує наявності достатнього обсягу пам’яті, що може бути критичним. Для перевірки необхідного обсягу пам’яті для зберігання важливої інформації треба оцінити один запис у базі. Для визначення необхідної кількості випробувань (у нашому випадку термін «кількість випробувань» може бути інтерпретовано як кількість текстів), результати яких застосовані для розрахунку коефіцієнтів precision, recall, forged, fallout та error, використано засоби математичної статистики та теорії ймовірності.

Сканування веб-сторінок є першим кроком в процесі збору даних. Елементи, що підлягають обробці, можуть відрізнятися в залежності від джерела даних. Створено специфікацію вимог до програмного забезпечення, що дозволяє у подальшій роботі над проектом чітко розуміти вимоги та обмеження до реалізації. Розроблено еталонну архітектуру та запропоновано варіант розгортання програмної системи (рис. 5). Розроблено програмні компоненти серверної частини програмної системи, що дозволяє проводити екстракцію даних з торгівельних площадок на основі використання гнучкого конфігурування та предикатної моделі видобування даних. Розроблено та імплементовано програмні компоненти для збору даних на прикладі збору характеристик моделей мобільних телефонів. Проведено тестування розроблених компонентів та доведено їх працездатність для збору даних з трьох різних торгівельних площадок.

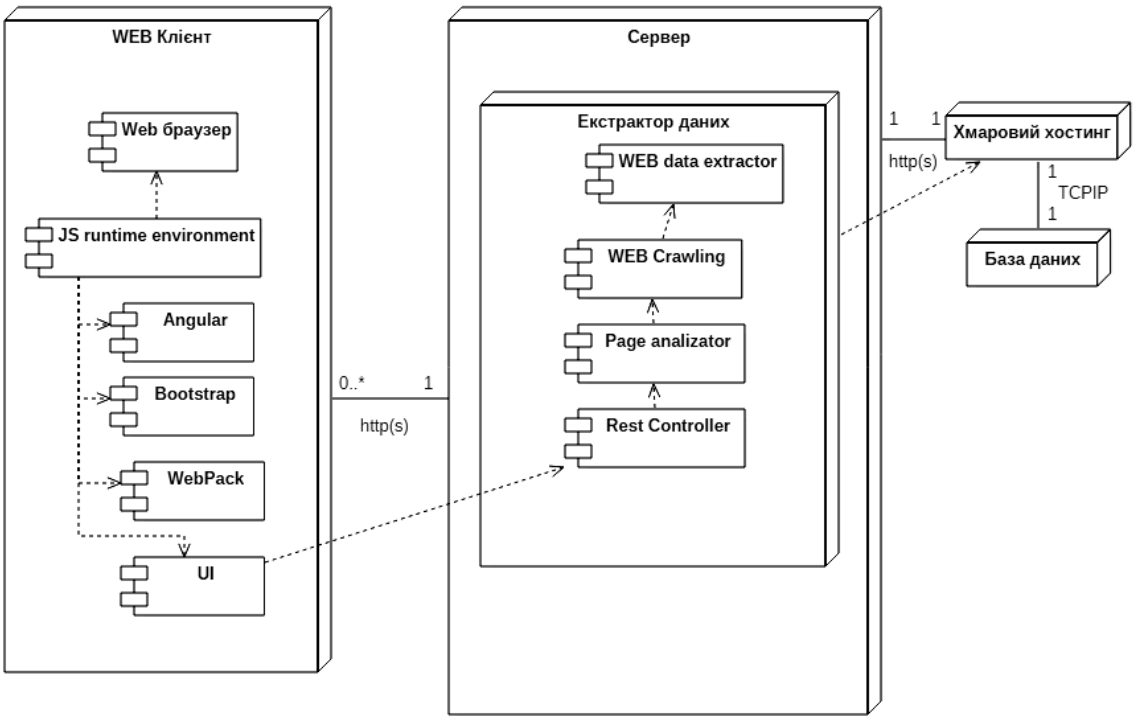

Рисунок 5 – Архітектура програмної системи збору та екстракції фактографічної інформації

Реалізацію запропонованої архітектури наведено у вигляді діаграми класів на рис. 6. Діаграма презентує основні класи, які розроблено в даному проекті: парсер, екстрактор та конфігуратор. Для досягнення результату у компонент подається URL необхідного джерела, що оброблюється у UrlUtils. Якщо джерело є коректним, то далі компонент завантажує всю HTML сторінку джерела. За дану дію відповідає клас Parser. Після отримання компонентом веб-сторінки він починає екстракцію даних, за яку відповідають класи Extractor та Configuration. Екстрактор використовує екземпляр класу Configuration для своїх налаштунків.

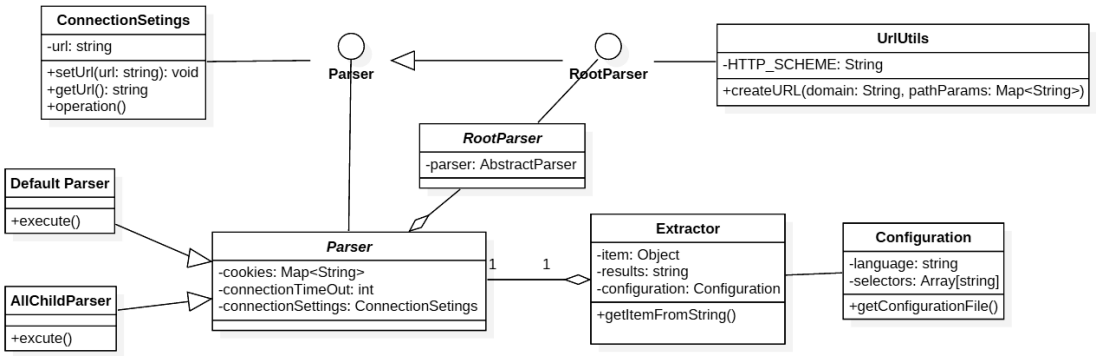

Рисунок 6 – Діаграма класів

Технологія фактографічного пошуку заснована на представленні змісту тексту у формі семантичної мережі, яка містить значимі слова і словосполучення, які зв’язані різними типами синтактико-семантичних зв’язків. Елементарна семантична мережа представляє результат синтаксичного аналізу та постсинтаксичних трансформацій дерева залежностей між словами у окремих реченнях. Повна семантична мережа тексту є сукупністю окремих семантичних мереж, які відповідають реченням.

У розділі описано застосування напівавтоматичного методу розширення базової онтології для предметної області «радіаційна безпека». Для вирішення проблеми неоднозначності слів використано словник синонімів. В розділі наводяться результати експерименту, виконаного для поповнення онтології новими екземплярами, знайденими в спеціалізованому текстовому корпусі. Радіаційна безпека – це сукупність засобів, які регулюються Міжнародними базовими нормами з радіаційної безпеки. Ці норми засновані на принципах Міжнародної комісії з радіаційної безпеки, основним завданням якої є зменшення ризику негативних наслідків для здоров’я внаслідок опромінювання радіацією або радіоактивними речовинами, які використовуються для промислових (вироблення енергії), медичних, сільськогосподарських, освітніх або дослідницьких цілей. Запропонована методика використана для розширення списку концептів базової онтології (БО) та встановлення між ними семантичних відношень синонімії. БО була визначена за допомогою експертів предметної області. У роботі представлені результати, які отримані на основі обробки понад 350 текстів російською мовою.

Відношення між будь-якими об’єктами можуть бути виявлені через порівняння їх властивостей. Властивість – це деякий атрибут, якість або характеристика об’єкта. У російській та українській мовах дієслова відіграють ключову роль в описі дій та інших властивостей суб’єкта. Іншими словами, дієслова явним чином виражають властивості суб’єкта. Слід відзначити, що у будь-якій мові існує множина лексичних способів вираження властивостей об’єктів. Наприклад, у двох фразах «*негативные эффекты радиационного облучения*» та «*радиационное облучение вызывает негативные эффекты*» для концепта «*радиационное облучение*» визначена властивість «*негативные эффекты*». Однак у першому випадку це здійснюється за допомогою родового відмінку, а у другому випадку – за допомогою дієслова «*вызывать»* у теперішньому часі. Таке різноманіття лінгвістичних форм істотно ускладнює процес обробки текстів природної мови. Для опису властивостей концептів були використані дієслова.

Для досягнення поставленої мети і розширення базової онтології предметної області новими екземплярами застосовано метод, оснований на використанні зовнішнього лінгвістичного ресурсу, а саме машинного словника синонімів (MRD). Для відбору релевантних термінів-кандидатів і для уникнення проблеми неоднозначності адаптовано метод аналізу формальних концептів – FCA. Ідея полягає у тому, щоб знайти у текстовому корпусі усі можливі лексичні входження даного концепту за допомогою зовнішнього лінгвістичного ресурсу, такого як словник синонімів, так як різні лексичні входження одного концепту зв’язані семантично. Для застосування методу необхідна реалізація декількох попередніх кроків: 1) морфологічна розмітка корпусу; 2) поверхневий аналіз речень; 3) перевірка і підтвердження предикатів; 4) підтвердження синонімів і маркування концептів різними лексичними варіантами. Суто процес поповнення онтології є заключним етапом.

У експерименті були використані тексти у лематизованій формі. Вихідними даними цього кроку є: $W = \{w\}$ –множина усіх іменників корпусу та $V = \{v\}$ – множина усіх дієслів корпусу без дублікатів. Завданням поверхневого аналізу речень є виявлення іменних груп, дієслівних груп та їх взаємних розташувань у реченні. Для цього були проіндексовані кожний іменник і дієслово. Припущення, що у російській мові більшість речень, написаних у академічному стилі, мають лінійну структуру з прямим порядком слів: підмет – підсудок – додаток, дозволило записати відношення між кожним суб’єктом (підметом) і його предикатом (підсудком) у вигляді пар (w, v). З кожного речення було збережено пари (w_i, v_j), де індекс іменника менше індексу дієслова $(i < j)$. Таким чином, встановлюються бінарні відношення між іменниками та дієсловами. Щоб відібрати предикати, необхідно визначити критерій їх відбору, для чого було розраховано вагу кожного дієслова. Для підтвердження синонімів і маркування концептів у якості вихідних даних використовується список концептів верхнього рівня $L = \{(l_i | i = 1, \dots n, n \in N)\}$, де l_i – лексичний варіант концепту в списку концептів. Із зовнішнього словника вилучається множина списків си-

нонімів $DL = \{DL_i | i = 1, \dots, n, n \in N\}$, які знайдено у ньому для кожного концепту; інакше: кожний DL_i є множиною синонімів із словника для концепту l_i. Перетин усіх списків синонімів і множини іменників корпусу дає список термінів-кандидатів: $CL_i = W \cup DL_i$ і $CL_i = \{w \subseteq W | \forall w: (w, l_i) \in I_{syn}\}$, де $I_{syn} \subseteq DL \times W$ означає відношення синонімії між іменниками корпусу та словами зі словника.

Для реалізації експерименту застосовано 9 основних концептів верхнього рівня: L ={*безопасность, защита, излучение, источник, население, облучение, персонал, риск, ущерб*}. Спочатку був створений спеціалізований текстовий корпус. Він складається з 58 галузевих стандартів, норм і звітів, які затверджено Міжнародною комісією з радіаційної безпеки і Національною комісією з радіаційного захисту України.

У теперішній час корпус складається більш ніж з 600 000 слів. Додатково були протестовані декілька тезаурусів і словників синонімів для російської мови, доступних on-line. Слід відзначити, що їх якість істотно відрізняється. У даному випадку був використаний Повний словник синонімів російської мови. Список синонімів для деяких понять предметної області представлено у табл. 1.

Для концепту «*ущерб*» у корпусі було виявлено лише 7 слів із 22 синонімів у словнику: $CL_{ущерб}$ = {*вред, потеря, повреждение, авария, осложнение, ухудшение, убыль*}. У корпусі виявлено 13 різних дієслів, які асоціюються з концептом «*ущерб*»: V={*включать, возмещать, вызывать, использовать, наносить, обеспечивать, ограничивать, оказаться, превышать, предотвращать, причинять, связывать, сопровождаться*}, однак після зважування лише 9 з них були відібрані у якості предикатів: {*включать, возмещать, вызывать, наносить, ограничивать, превышать, предотвращать, причинять, сопровождаться*}. Результат реалізації цього процесу показаний на рис. 7.

Таблиця 1 – Список синонімів для визначення концептів предметної області

Поняття	Синоніми з зовнішнього словника
Ущерб	*убыток, потери, вред, урон, изъян, потеря, повреждение, поломка, авария, протори и убытки, подрыв, осложнение, утрата, разор, ухудшение, протори, порча, невыгода, пагуба, шкода, брешь, наклад*
Излучение	*изливание, излитие, источение, свет, испускание, эманация, радиация, лучеиспускание, сноп, фонирование*

Нарешті, у відповідності з визначенням формального концепту, який подано вище, лише 4 кандидати були відібрані та 3 з них підтверджені експертом. Для поняття «*излучение*» у корпусі був знайдений тільки один синонім, і він також був підтверджений за допомогою FCA й схвалений експертом. Ці результати представлено в табл. 2.

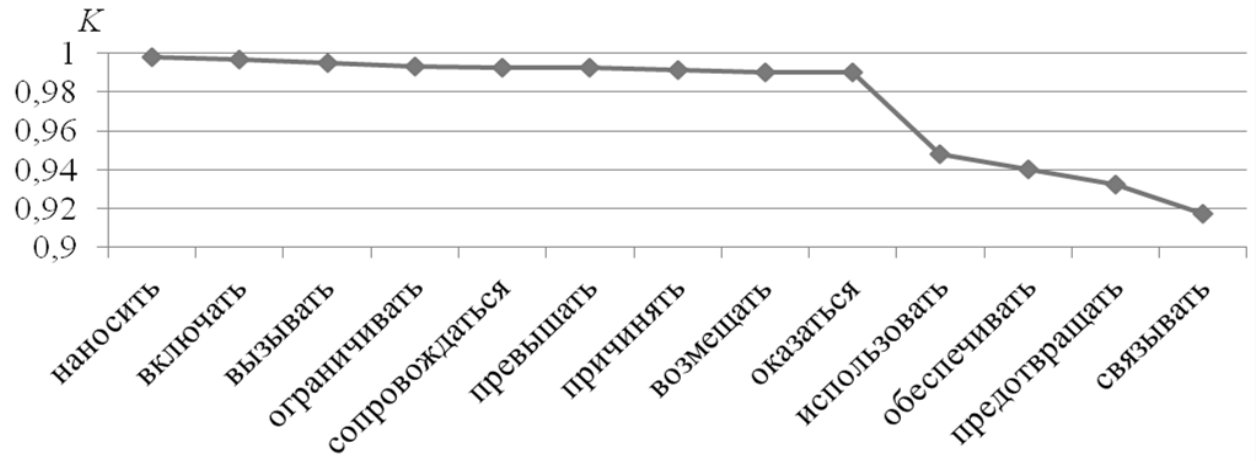

Рисунок 7 – Відбір предикатів (по вертикальній осі відкладені значення узагальненого коефіцієнта зважування *K*)

Таблиця 2 – Список підтверджених синонімів для розширення базової онтології

Поняття базової онтології	Список синонімів
Ущерб	*вред, повреждение, авария, осложнение*
Излучение	*радиация*

Релевантність синонімів, відібраних для інших концептів, склала від 50 до 100 % . Аналогічний експеримент був проведений для паралельного французького корпусу з використанням електронного словника синонімів CRISCO. У результаті експерименту визначено, що якість отриманих результатів для французької версії онтології виявилась вищою, що можна пояснити недостатньою повнотою російськомовного ресурсу та більш складною структурою речень статей корпусу, що потребує додаткового, більш тонкого аналізу.

У цьому ж розділі представлено оцінку ефективності та перспективи використання отриманих моделей та методів. Оцінку ефективності здійснено окремо для двох основних задач, які вирішуються у дослідженні: задачі видобування фактів з текстів та задачі видобування фактів разом з їх визначеннями. Для оцінки ефективності використані коефіцієнт точності *Precision* та коефіцієнт повноти *Recall*. Обрані коефіцієнти розраховуються на основі показників: *a* – кількість видобутих текстів-кандидатів, які є фактами; *b* – кількість видобутих текстів-кандидатів, які не є фактами; *c* – кількість не видобутих текстів-кандидатів, які є фактами; *d* – кількість не видобутих текстів-кандидатів, які не є фактами. Проаналізовано близько 400 документів предметних областей «радіаційна безпека» та «обробка патентно-кон'юнктурної інформації». Для задачі видобування фактів з текстів отримано наступні значення: *Recall* = 0,89, *Precision* = 0,94. Для задачі видобування фактів разом з їх визначеннями *Recall* = 0,67, *Precision* = 0,89. Порівняння отриманих результатів з результата-

ми подібних систем показали ефективність роботи інформаційної системи щодо вирішення представлених двох задач.

Показано перспективи використання запропонованих моделей і методів ідентифікації та обробки предметних знань для індексування повнотекстових документів у задачах інтелектуального пошуку фактографічної інформації за ключовими словами, розробки інформаційної технології створення OLAP-кубів для подання багатовимірного простору знань колекції документів, визначення семантичної близькості на основі когнітивного підходу.

Таким чином, розроблена інформаційна технологія інтелектуального аналізу фактографічної інформації удосконалює та доповнює існуючий підхід обробки текстових даних і не суперечить існуючій практиці, що свідчить про її практичну цінність та ефективність використання.

ВИСНОВКИ

У дисертаційній роботі вирішено актуальну науково-практичну задачу розробки моделей та інформаційної технології інтелектуального аналізу фактографічних ресурсів для забезпечення несуперечності та актуальності результатів інтелектуального аналізу фактографічних ресурсів.

У ході виконання дисертаційної роботи отримані наступні результати.

1. Проаналізовано існуючі інформаційні технології, моделі та методи обробки фактографічних даних у мережевих потоках слабоструктурованої текстової інформації, сформульовано основні вимоги до розробки інформаційної технології інтелектуального аналізу фактографічних ресурсів.

2. У якості математичного інструментарію моделювання фактів визначено теорію категорій, у тому числі її проективну та предикатну інтерпретації. Запропоновано для аналізу фактографічної інформації використовувати метод екстракції фактографічних даних, заснований на методі компараторної ідентифікації, що дозволяє вирішити задачу моделювання інформаційного пошуку на основі семантичного оцінювання триплетів фактів.

3. Запропоновано розв’язання задачі інтелектуального аналізу мережевих текстових фактографічних ресурсів на основі використання інструментів алгебри скінченних предикатів і методів теорії інтелекту, що дозволило скоротити час та збільшити точність тематичного інформаційного пошуку за смисловими характеристиками фактів.

4. Розроблено моделі тематичного пошуку та екстракції фактографічної інформації на основі інтелектуальної процедури оцінки текстової інформації. Запропоновано для опису фактів використання двох типів триплетів: «Суб’єкт $\rightarrow$ Предикат $\rightarrow$ Об’єкт» та «Предмет $\rightarrow$ Атрибут $\rightarrow$ Значення», що дозволяє вилучати поняття зі слабо структурованих текстових ресурсів і описувати відношення між ними у структурованому вигляді.

Сформовано підхід до видобування фактографічних даних із текстових джерел та їх інтеграції на основі використання онтологій. Запропоновано вико-

ристання нового напівавтоматичного методу для розширення базової предметної онтології на прикладі предметних областей «радіаційна безпека» та «обробка ПКІ». Запропонований підхід до автоматизованої побудови онтології дозволяє удосконалити та доповнити існуючий підхід обробки текстових даних, що свідчить про його практичну цінність та ефективність використання.

5. Розроблено інформаційну технологію інтелектуального аналізу текстової інформації, яка дозволяє автоматизувати процес обробки слабоструктурованих фактографічних ресурсів та вдосконалити процес екстракції знань за рахунок визначення ознак та атрибутів концептів предметних областей.

6. Проведено апробацію розроблених моделей, підходів та інформаційної технології та впроваджено результати дослідження в існуючі інформаційні системи. Створено специфікацію вимог до програмного забезпечення, розроблено еталонну архітектуру та запропоновано варіант розгортання програмної системи. Розроблено програмні компоненти серверної частини програмної системи, що дозволяє проводити екстракцію даних на основі використання гнучкого конфігурування та предикатної моделі видобування даних.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Дорошенко А. Ю. Построение онтологий и фреймворк информационной системы для создания интеллектуальной системы / А. Ю. Дорошенко, Е. А. Оробинская, О. И. Король // Вісник Херсонського національного технічного університету. – Херсон : ХНТУ, 2013. – № 1 (46). – С. 196–200.

Здобувачеві належить постановка задачі створення інтелектуального додатку, здатного виявляти у тексті релевантні відношення концептів.

2. Дорошенко А. Ю. Применение масштабных лингвистических ресурсов для расширения онтологии предметной области (на примере области «Радиационная безопасность») / Е. А. Оробинская, Н. В. Шаронова, А. Ю. Дорошенко, Ж.-Ю. Шоша // Східно-Європейський журнал передових технологій. – 2014. – № 5/2 (71). – С. 9–14.

Здобувачем досліджено базові засади процесу обробки лінгвістичних ресурсів для розширення онтології предметної області.

3. Дорошенко А. Ю. Интеллектуальные технологии идентификации фактографической информации / А. Ю. Дорошенко, Е. А. Оробинская, Аджит Пратап Сингх Гаутам // Проблеми інформаційних технологій. – Херсон : ХНТУ, 2014. – № 2 (016). – С. 103–106.

Здобувачем запропоновано застосування методу компараторної ідентифікації для використання у системах обробки фактографічної інформації.

4. Дорошенко А. Ю. Розробка програмних компонентів інформаційної системи екстракції фактографічних даних з веб-ресурсів / А. Ю. Дорошенко, Н. В. Шаронова, Б. О. Єна, О. В. Янголенко // Проблеми інформаційних технологій. – Херсон : ХНТУ, 2018. – № 1 (023). – С. 27–35.

Здобувачем розроблено компонентну архітектуру підсистеми екстракції фактографічних даних та їх оцінювання.

5. Дорошенко А. Ю. Розробка інформаційної технології інтелектуального аналізу фактографічної інформації / А. Ю. Дорошенко // Біоніка інтелекту. – Харків : ХНУРЕ, 2018. – № 1 (90). – С. 116–121.

6. Пат. на корисну модель 63508 Україна, МПК G06F 17/18. Цифровий гібридний медіанний фільтр / А. В. Шостак, А. Ю. Дорошенко, Ю. І. Дорошенко, М. Г. Коробков ; заявник та патентовласник Національний аерокосмічний університет «ХАІ». – № 201103302; заявл. 21.03.2011; опубл. 10.10.2011, Бюл. № 19. – 4 с.

Здобувачеві належить лінгвістичний опис корисної моделі цифрового медіанного фільтру.

7. Пат. на корисну модель 62818 Україна, МПК G06F 17/18. Пристрій цифрової фільтрації сигналу / А. В. Шостак, А. Ю. Дорошенко, Ю. І. Дорошенко, М. Г. Коробков, О. М. Рисований, А. В. Івашко ; заявник та патентовласник НТУ «ХПІ». – № 201105823; заявл. 10.05.2011; опубл. 12.09.2011, Бюл. № 17. – 9 с.

Здобувач здійснила порівняльний аналіз відомих рекурсивних медіанних фільтрів.

8. Дорошенко А. Ю. Формальна модель природної мови як важлива частина прогресивних інформаційних технологій / А. Ю. Дорошенко, Т. О. Богдан // Proceedings of the III International Conference on Computer Science and Information Technologies. – Lviv : Institute of Computer Science and Information Technologies, 2008. – P. 394–396.

Здобувачем виконано аналіз інформаційних технологій з точки зору застосування в них моделей природної мови.

9. Дорошенко А.Ю. Использование онтологий для автоматической обработки текстов на естественном языке / А. Ю. Дорошенко, Е. А. Оробинская // Вісник Національного технічного університету «Харківський політехнічний інститут»: Тематичний вип. : Актуальні проблеми розвитку українського суспільства. – Харків : НТУ «ХПІ», 2011. – № 30. – С. 101–106.

Здобувачем розроблено підхід до використання онтологій для автоматизованої обробки природномовних текстів.

10. Doroshenko A. Y. The problem of studying foreign languages in technical universities / A. Y. Doroshenko, A. V. Kovalyova // Integration Processes and Innovative Technologies: Achievements and Prospects of Engineering Sciences. – Kharkiv : Kharkiv National Automobile and Highway University, 2011. – P. 272–275.

Здобувачем розглянуто проблеми вивчення іноземної мови у технічних університетах з точки зору впровадження інноваційних технологій.

11. Дорошенко А. Ю. Використання онтологій для семантичного пошуку документів / А. Ю. Дорошенко // Інтелектуальні системи та прикладна лінгвіс-

тика : тези доп. Першої Всеукраїнської науково-практ. конф. – Харків : НТУ «ХПІ», 2012. – С. 8–9.

12. Дорошенко А. Ю. Використання онтологій для семантичного пошуку документів / А. Ю. Дорошенко, Сайед Мохаммад Таухід Сіддікі, Н. В. Шаронова // Вісник Національного технічного університету «Харківський політехнічний інститут». Тематичний вип. : Актуальні проблеми розвитку українського суспільства. – Харків : НТУ «ХПІ», 2012. – № 31. – С. 95–99.

Здобувач запропонувала спосіб використання онтологій для семантичного пошуку документів.

13. Дорошенко А. Ю. Системы обработки патентно-конъюнктурной информации на основе онтологий / Н. В. Шаронова, А. Ю. Дорошенко // Інформаційні проблеми теорії акустичних, радіоелектронних і телекомунікаційних систем : тези доп. другої Міжнар. наук.-техн. конф. – Харків : НТУ «ХПІ», 2013. – С. 37–38.

Здобувачеві належить постановка задачі, формування системи ознак побудови онтології для обробки патентно-кон'юнктурної інформації.

14. Дорошенко А. Ю. Обработка фактографической информации для текстовых патентно-конъюнктурных данных при построении онтологий / А. Ю. Дорошенко // Експертні оцінки елементів навчального процесу : матеріали XV міжвуз. наук.-практ. конф. – Харків : вид-во НУА, 2013. – С. 29–31.

15. Дорошенко А. Ю. Система построения онтологий для обработки фактографической информации на примере текстовых патентно-конъюнктурных данных / А. Ю. Дорошенко // Актуальні проблеми прикладної лінгвістики очима наукової молоді : матеріали III регіональної науково-практ. конф. – Харків : НАУ ім. М. Є. Жуковського «ХАІ», 2013. – С. 36–37.

16. Дорошенко А. Ю. Извлечение информации из текстовых сообщений на основе правил EBNF [Електронний ресурс] / О. В. Канищева, А. Ю. Дорошенко // Прикладна лінгвістика та лінгвістичні технології Megaling-2013 : зб. наук. пр. – Режим доступу: http://megaling.ulif.org.ua/tezi-2013-rik/storinka-13.html.

Здобувач розробила засоби екстракції фактографічної інформації з потоку текстових повідомлень.

17. Doroshenko A. Issues of Fact-based Information Analysis [Electronic resource] / N. Sharonova, A. Doroshenko, O. Cherednichenko // Proceedings of the International Conference on Computational Linguistics and Intelligent Systems. – 2018. – URL: http://ceur-ws.org/Vol-2136/10000011.pdf. (індексовано в Scopus).

Здобувачем досліджено базові засади процесу екстракції фактографічної інформації для уникнення невідповідностей при описі товарів.

18. Doroshenko A. Towards the Ontology-Based Approach for Factual Information Matching / N. Sharonova, A. Doroshenko, O. Cherednichenko // Информационные системы и технологии (ИСТ-2018) : Інформаційні системи і технології (ІСТ-2018) : матеріали VII Міжнарод. наук.-техн. конф. – Харків : ХНУРЕ, 2018. – С. 230–233.

Здобувачем досліджено процес екстракції фактографічної інформації для зняття протиріч при описі товарів.

АНОТАЦІЇ

Дорошенко А. Ю. Інформаційна технологія інтелектуального аналізу фактографічних текстових ресурсів. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Національний технічний університет «Харківський політехнічний інститут», Харків, 2019.

У дисертаційній роботі вирішена актуальна науково-практична задача розробки моделей та інформаційної технології інтелектуального аналізу фактографічної інформації. На основі аналізу моделей та методів обробки фактографічних даних у мережевих потоках сформульовано основні вимоги до розробки інформаційної технології інтелектуального аналізу фактографічних ресурсів. У якості математичного інструментарію моделювання фактів визначено теорію категорій, її проективну та предикатну інтерпретації. Запропоновано для опису фактографічної інформації використовувати теорію інтелекту, метод компараторної ідентифікації та апарат алгебро-логічних рівнянь. Розроблено моделі тематичного пошуку та екстракції фактографічної інформації на основі інтелектуальної процедури оцінки текстової інформації. Запропоновано для опису фактів використання двох типів триплетів: «Суб'єкт → Предикат → Об'єкт» та «Предмет → Атрибут → Значення», що дозволяє вилучати поняття зі слабоструктурованих текстових ресурсів та описувати відношення між ними у структурованому вигляді. Сформовано підхід до видобування фактографічних даних з текстових джерел, запропоновано використання онтологій для опису процесів інтеграції фактографічної інформації. Запропоновано використання нового напівавтоматичного методу для розширення базової онтології на прикладі предметних областей «радіаційна безпека» та «обробка патентно-кон'юнктурної інформації». Проведено апробацію розроблених моделей, підходів та інформаційної технології та впроваджено результати дослідження у реальні інформаційні системи. Розроблено еталонну архітектуру, програмні компоненти серверної частини програмної системи, що дозволяє проводити екстракцію даних на основі використання гнучкого конфігурування та предикатної моделі видобування даних.

Ключові слова: інформаційна технологія, фактографічна інформація, метод компараторної ідентифікації, екстракція фактів, онтологічна специфікація.

Дорошенко А. Ю. Информационная технология интеллектуального анализа фактографических текстовых ресурсов. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 – информационные технологии. – Национальный

технический университет «Харьковский политехнический институт», Харьков, 2019.

В диссертационной работе решена актуальная научно-практическая задача разработки моделей и информационной технологии интеллектуального анализа фактографических текстовых ресурсов. Проанализированы существующие информационные технологии, модели и методы обработки фактографических данных в сетевых потоках слабоструктурированной текстовой информации, сформулированы основные требования к разработке информационной технологии интеллектуального анализа фактографических ресурсов. В качестве математического инструментария моделирования фактов определена теория категорий, в том числе ее проективная и предикатная интерпретации. Рассмотрена возможность использования предикатных категорий для формализации знаний, представленных с помощью логических моделей. Для описания фактографической информации предложено использовать средства теории интеллекта, метод компараторной идентификации и аппарат алгебро-логических уравнений.

Разработана логико-лингвистическая модель информационного поиска фактографической информации на основе оценивания веб-страницы с применением компаратора. Разработаны модели тематического поиска и экстракции фактографической информации на основе интеллектуальной процедуры оценки текстовой информации. Предложено для описания фактов использование двух типов триплетов: «Субъект → Предикат → Объект» и «Предмет → Атрибут → Значение», что позволяет извлекать понятия из слабоструктурированных текстовых ресурсов. Сформирован подход к извлечению фактографических данных из текстовых источников на основе использования онтологий. Описано использование онтологий для описания процессов интеграции фактографической информации. Предложено использование нового полуавтоматического метода для расширения базовой онтологии, на примере предметных областей «радиационная безопасность» и «обработка патентно-конъюнктурной информации». Разработана комплексная информационная технология интеллектуального анализа фактографической информации, которая позволяет автоматизировать процесс обработки слабоструктурированных фактографических ресурсов и усовершенствовать процесс экстракции знаний за счет определения признаков и атрибутов концептов предметных областей.

Проанализированы подходы и информационные технологии решения задач парсинга на базе существующих информационных систем. Создана спецификация требований к программному обеспечению. Разработана эталонная архитектура и предложен вариант развертывания программной системы. Разработаны программные компоненты серверной части программной системы, что позволяет проводить экстракцию данных на основе использования гибкого конфигурирования и предикатной модели извлечения данных.

Результаты исследования внедрены в практику построения подсистем поиска фактографической информации в научных библиотеках вузов Харькова, а

также в учебный процесс кафедры интеллектуальных компьютерных систем НТУ «ХПИ».

Ключевые слова: информационная технология, фактографическая информация, метод компараторной идентификации, экстракция фактов, онтологическая спецификация.

Doroshenko A. Y. Information Technology of Intellectual Analysis of the Fact-based Text Resourses. – Manuscript.

The dissertation for a candidate degree in technical sciences, specialty 05.13.06 – Information Technologies. – National Technical University «Kharkiv Polytechnic Institute», Kharkiv, 2019.

The actual scientific and practical task of developing models and information technology of intellectual analysis of factual information is solved in the dissertation. On the basis of analysis of models and methods of processing factual data in network streams, the basic requirements for the development of information technology of intellectual analysis of factual resources are formulated. The theory of categories, its projective and predicate interpretations is determined as a mathematical tool for modeling facts. It is proposed to use the theory of intelligence, the method of comparative identification and the apparatus of algebra-logical equations to describe factual information. Models of thematic search and extraction of factual information on the basis of the intellectual procedure for evaluating textual information have been developed. It is proposed to describe the use of two types of triplets: "Subject → Predicate → Object" and "Item → Attribute → Value", which allows you to remove the concept of weakly structured text resources and describe the relationship between them in a structured form. An approach to extracting factual data from text sources has been formed, and the use of ontologies for the description of the processes of integration of factual information is proposed. The use of a new semi-automatic method is proposed for extending the basic ontology, on the example of the subject areas "radiation safety" and "processing of patent information". Approbation of developed models, approaches and information technology was carried out and the results of research were implemented in real information systems. The reference architecture, software components of the server part of the software system, which allows data extraction based on the use of flexible configuration and predicate data mining model, is developed.

Keywords: information technology, factual information, comparative identification method, fact extraction, ontological specification.

Підп. до друку 04.03.2019 р. Формат 60х90/16.
Гарнітура Times New Roman. Папір офсетний.
Друк – цифровий. Ум. друк. аркушів 0,9
Наклад 100 прим. Зам. №

Надруковано у ФЛ-П Черняк Л.О.
61002, м. Харків, вул. Багалія, 16
Свідоцтво № 24800000000079553, від 16.05.2007 р.