

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

ПЕТРАСОВА СВІТЛАНА ВАЛЕНТИНІВНА



УДК 004.912:007.51

**ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ
ІДЕНТИФІКАЦІЇ ЗНАНЬ У НАУКОМЕТРИЧНИХ СИСТЕМАХ
НА ОСНОВІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ
СЛАБОФОРМАЛІЗОВАНИХ ДАНИХ**

Спеціальність 05.13.06 – інформаційні технології

Автореферат дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2017

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
Хайрова Ніна Феліксівна,
Національний технічний університет
«Харківський політехнічний інститут», професор
кафедри інтелектуальних комп'ютерних систем

Офіційні опоненти: доктор технічних наук, професор
Шостак Ігор Володимирович,
Національний аерокосмічний університет
ім. М. Є. Жуковського «Харківський авіаційний
інститут», професор кафедри інженерії програмного
забезпечення

кандидат технічних наук, доцент
Чала Лариса Ернестівна,
Харківський національний університет радіо-
електроніки, доцент кафедри штучного інтелекту

Захист відбудеться «27» квітня 2017 р. о 15.00 годині 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 «___» _____ 2017 р.

Вчений секретар
спеціалізованої вченої ради



Дорофєєв Ю.І.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. В останні роки в інформаційній практиці спостерігається зростаючий інтерес до наукометричних досліджень. При цьому у переважній більшості випадків накопичені гігабайти бібліографічної та реферативної інформації наукометричних систем вимагають якісно нових форм аналітико-синтетичної обробки.

Інформаційні простори, що позначають ці ресурси та представляють фронти наукових досліджень, утворені явно вираженими відношеннями (цитуванням, авторством, ключовими словами, спільністю теми), які отримані статистичними або імовірнісними методами, у переважній більшості представляються недостатніми і втрачають частину знань про спільність інформації.

Складність аналізу слабоформалізованих рефератів наукометричних систем визначається багатозначністю і синонімічністю, які властиві мові на всіх рівнях її представлення, що, перш за все, впливає на визначення смислової єдності мовних одиниць. При цьому проблемність завдання підвищується експоненціально, якщо йдеться про смислову близькість великих інформаційних фрагментів.

У зв'язку зі стрімким зростанням обсягів неструктурованих інформаційних ресурсів у наукометричних системах та існуючими підходами і методами аналізу слабоформалізованих даних, використання яких досягло межі своєї реалізації, перспективності набувають завдання обробки слабоформалізованої інформації на базі інтелектуального аналізу. Інтелектуальним аналізом слабоформалізованої інформації займалися Ю.Д. Апресян, М.Ф. Бондаренко, Т.А. Виноград, Дж.А. Клейнберг, Д.В. Ланде, К.Д. Маннінг, Ч.Дж. Філлмор, Н.Ф. Хайрова, А.Н. Хомський, Ю.П. Шабанов-Кушнарєнко, Н.В. Шаронова, Р.К. Шенк, В.А. Широков та інші. Враховуючи особливості ідентифікації знань у слабоформалізованій інформації, виникає доцільність подальшого удосконалення обробки природномовної інформації в реферативних наукометричних системах.

Таким чином, науково-практична задача підвищення ефективності процесу ідентифікації знань у наукометричних системах за рахунок використання інформаційно-логічних моделей і методів інтелектуального аналізу слабоформалізованої інформації 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), а також госпдоговірної теми «Дослідження засобів інтелектуального аналізу тексту і алгоритмів семантичного пошуку для текстів економічної проблематики» з компанією «Document Concept AG» (Обервіль, Швейцарія, рамковий договір № 03202 від 31.11.2011 р.), у яких здобувач брала участь як виконавець.

Мета і задачі дослідження. Метою дисертаційної роботи є підвищення ефективності процесу ідентифікації знань у наукометричних системах за рахунок побудови моделей і методів інтелектуального аналізу слабоформалізованих даних.

Відповідно до зазначеної мети поставлені задачі:

- проаналізувати сучасний стан проблеми виявлення знань у наукометричних системах, систематизувати існуючі засоби інтелектуального аналізу слабоформалізованої інформації та сформулювати основні вимоги до розробки інформаційної технології ідентифікації знань;
- дослідити можливість використання інструментарію алгебри скінченних предикатів для побудови інформаційно-логічних моделей ідентифікації знань у наукометричних системах;
- побудувати модель визначення семантично зв'язних фрагментів слабоформалізованої реферативної інформації;
- розробити метод формалізації смислових відношень сутностей в реферативних текстових потоках на основі інтелектуального аналізу;
- реалізувати метод компараторної ідентифікації смислових фрагментів рефератів наукометричних систем;
- удосконалити інформаційну технологію ідентифікації знань у слабоформалізованій інформації наукометричних систем;
- здійснити практичну реалізацію розробленої інформаційної технології, визначити оцінку ефективності розробленої системи і впровадити отримані наукові результати дисертаційної роботи в практику.

Об'єктом дослідження є процес ідентифікації знань у слабоформалізованій інформації наукометричних систем.

Предметом дослідження є моделі та методи ідентифікації знань у наукометричних системах.

Методи досліджень. Теоретичною базою виконаних досліджень є фундаментальні положення теорії інтелекту та алгебри скінченних предикатів. Алгебра скінченних предикатів застосована для опису семантико-граматичних характеристик сутностей при створенні інформаційно-логічної моделі ідентифікації семантично зв'язних сутностей. Метод компонентного аналізу та інструментарій регулярних виразів використовуються для формалізації смислових відношень толерантності та еквівалентності у слабоформалізованій реферативній інформації. Метод компараторної ідентифікації використаний для опису процесу формування єдиних інформаційних просторів наукової взаємодії авторів рефератів у наукометричних системах. Для підтвердження достовірності отриманих результатів використано методи математичної статистики.

Наукова новизна отриманих результатів:

- *уперше* розроблено логіко-лінгвістичну модель визначення семантично зв'язних фрагментів реферативної інформації, яка відрізняється застосуванням алгебро-предикатних операцій для формалізації семантико-граматичних характеристик сутностей та дозволяє вдосконалити інформаційну технологію ідентифікації знань у слабоформалізованих даних;

– *отримав подальший розвиток* метод компараторної ідентифікації смислових фрагментів рефератів, що дозволяє виділити єдині інформаційні простори наукової взаємодії авторів за рахунок моделювання функцій інтелекту з розуміння та класифікації смислу;

– *удосконалено* метод формалізації смислових відношень сутностей, який базується на використанні міри смислової близькості і відрізняється застосуванням інтелектуального аналізу при виявленні класів еквівалентності та толерантності, що дозволяє визначити неявно виражені відношення близькості й таксономії;

– *удосконалено* інформаційну технологію ідентифікації знань у наукометричних системах, яка відрізняється етапами визначення імпліцитних зв'язків між рефератами наукометричних систем, що дозволяє динамічно виявляти спільні фронти наукових досліджень.

Практичне значення одержаних результатів для удосконалення наукометричних інформаційних систем полягає у розробці технології ідентифікації єдиних інформаційних просторів наукової взаємодії авторів. Розроблена технологія включає логіко-лінгвістичну модель ідентифікації семантично зв'язних фрагментів слабоформалізованої реферативної інформації, метод формалізації смислових відношень толерантності та еквівалентності сутностей, метод компараторної ідентифікації смислових фрагментів рефератів наукометричних систем. Основні практичні результати роботи можуть бути використані в інформаційно-пошукових, експертних, інформаційно-аналітичних системах широкого призначення, при вирішенні завдань формування електронних каталогів зв'язних за змістом текстів у наукометричних, бібліотечних та реферативних системах.

Результати дослідження знайшли практичне застосування у системах обробки анотацій та рефератів у Науковій бібліотеці Харківського національного університету радіоелектроніки, Науково-технічній бібліотеці НТУ «ХП», Науковій бібліотеці Національної юридичної академії України ім. Ярослава Мудрого. Використання розроблених у роботі моделей і методів дозволило підвищити ефективність процесу ідентифікації знань у слабоформалізованій реферативній інформації за рахунок підвищення середніх значень коефіцієнтів повноти й точності видачі спільної за змістом інформації.

Теоретичні результати дисертації використовуються у навчальному процесі на кафедрі інтелектуальних комп'ютерних систем НТУ «ХП» при викладанні спеціальних дисциплін «Сучасні інтелектуальні системи аналізу та обробки інформації», «Автоматизована обробка природної мови», «Штучний інтелект: лінгвістичні проблеми» для студентів спеціальності «Прикладна лінгвістика» та при виконанні курсових й дипломних робіт.

Особистий внесок здобувача. Усі основні результати дисертаційної роботи, що виносяться на захист, отримані здобувачем особисто. Серед них: аналіз основних підходів до ідентифікації знань у наукометричних системах та основних засобів інтелектуальної обробки слабоформалізованих даних; логіко-лінгвістична модель ідентифікації семантично зв'язних фрагментів слабоформалізованої реферативної інформації; метод формалізації смислових відношень

толерантності та еквівалентності; метод компараторної ідентифікації смислових фрагментів рефератів наукометричних систем; інформаційна технологія ідентифікації знань у наукометричних системах; метод експериментальної оцінки ефективності розробленої інформаційної технології ідентифікації знань.

Апробація результатів дисертації. Результати дисертаційної роботи доповідались та обговорювались на: Міжнародній науковій конференції MegaLing «Горизонти прикладної лінгвістики і лінгвістичних технологій» (Київ, 2013); XV, XVI міжвузівських науково-практичних конференціях «Експертні оцінки елементів учбового процесу» (Харків, 2013, 2014); III Регіональній науково-практичній конференції «Актуальні проблеми прикладної лінгвістики очима наукової молоді» (Харків, 2013); 18-му Міжнародному молодіжному форумі «Радіoeлектроніка та молодь у XXI столітті» (Харків, 2014); VII International Conference on Intelligent Information and Engineering Systems (Rzeszow, Poland, 2014); III, IV, V Всеукраїнських науково-практичних конференціях «Інтелектуальні системи та прикладна лінгвістика» (Харків, 2014, 2015, 2016); XXIII, XXIV Міжнародних науково-практичних конференціях «Інформаційні технології: наука, техніка, технологія, освіта, здоров'я» (Харків, 2015, 2016); X International Scientific and Technical Conference on Computer Science & Information Technologies – CSIT (Львів, 2015).

Публікації. Основні результати дисертації опубліковані у 19 наукових працях, серед яких 4 статті у фахових наукових виданнях України, 3 статті в іноземних періодичних фахових виданнях (1 – Scopus), 12 – у матеріалах конференцій (1 – Scopus).

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Повний обсяг дисертації складає 145 сторінок, з них 22 рисунки та 14 таблиць по тексту, 133 найменувань використаних джерел на 15 сторінках, 2 додатки на 14 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** на основі аналітичного огляду інформаційних джерел проаналізовано існуючі проблеми ідентифікації знань у наукометричних системах, які включають бібліографічну та реферативну інформацію, а також інструмент для відстеження цитування наукових статей. Відзначено, що при виявленні спільних фронтів наукових досліджень використання зв'язків цитування не можуть бути застосовані у якості універсального засобу, оскільки частина знань про спільність інформації, виражених імпліцитними зв'язками між рефератами наукометричних систем, втрачається.

Вирішення завдання ідентифікації знань ускладнюється проблемами полісемії, синонімії, омонімії природномовних джерел знань, які властиві мові на всіх рівнях її представлення (морфологічному, синтаксичному, семантичному і прагматичному), що, в першу чергу, проявляється в проблемі визначення відношення смислової еквівалентності.

Проведений аналітичний огляд існуючих технологій, підходів і методів інтелектуального аналізу слабоформалізованих даних (рис. 1) показав, що для реалізації автоматичної ідентифікації сутностей та відношень (зокрема, відношень еквівалентності та толерантності) стає перспективним застосування семантико-синтаксичних засобів аналізу у комплексі з моделями і методами когнітивної діяльності людини.

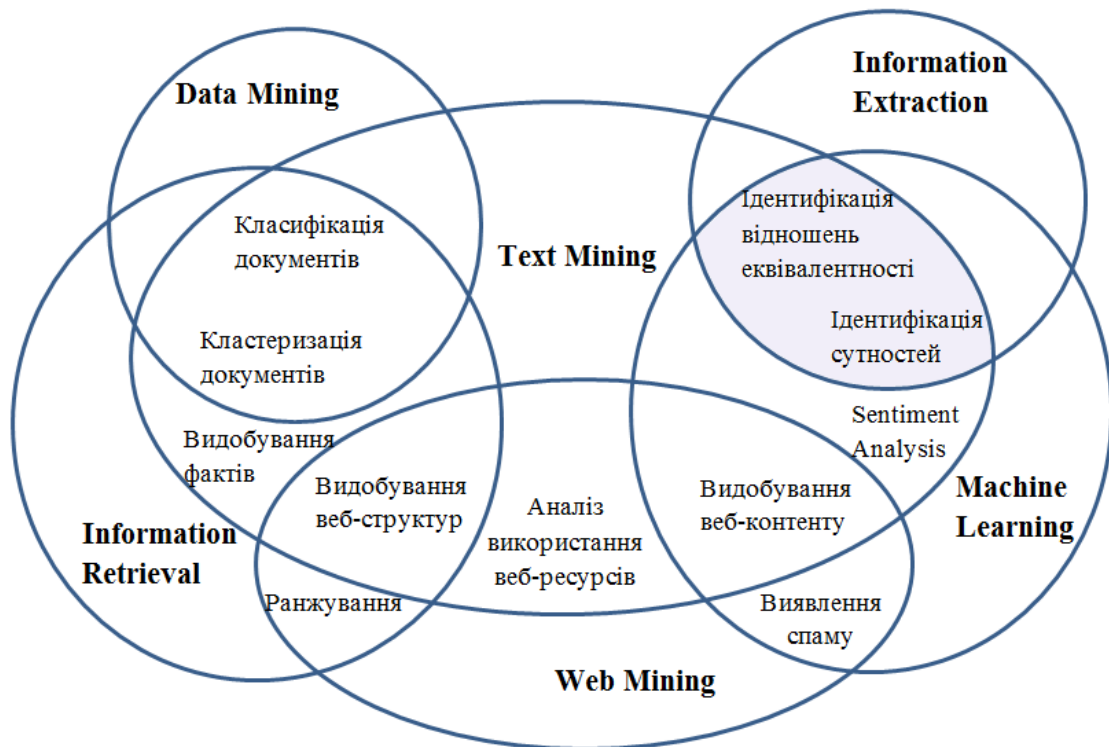


Рисунок 1 – Структурна схема технологій і завдань інтелектуального аналізу інформації

Сформульовано й обґрунтовано доцільність розробки моделей та методів інтелектуального аналізу слабоформалізованої інформації для підвищення рівня автоматизації обробки інформації та подальшого їх впровадження в бібліотечні системи при формуванні інформаційних просторів наукової взаємодії авторів.

У другому розділі дисертації обґрунтовано вибір математичних засобів ідентифікації та формалізації нових знань в реферативній інформації.

У якості математичного апарату побудови моделі ідентифікації дискретного скінченного набору смислових сутностей і відношень у слабоформалізованій реферативній інформації наукометричних систем використовується алгебра скінченних предикатів.

При математичному описі функцій інтелекту вказуються відповідні цим функціям алфавітні оператори, що перетворюють вхідні слова у вихідні.

Кожному скінченному алфавітному оператору F ставимо у відповідність деякий скінченний предикат P , який дорівнює 1, якщо скінченний оператор перетворює вхідні слова x довжини n у вихідні слова y тієї ж довжини, що складаються з літер алфавіту A . В іншому випадку скінченний предикат дорівнює 0.

Скінченний предикат над алфавітом A – це будь-яка функція $P(x_1, x_2, \dots, x_n)$ від n буквених аргументів $x_1, x_2, \dots, x_n$, які задані на множині A , що приймає логічне значення істини або хибності

$$P(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n) = \begin{cases} 1, & \text{якщо } F(x_1 x_2 \dots x_n) = y_1 y_2 \dots y_n, \\ 0, & \text{якщо } F(x_1 x_2 \dots x_n) \neq y_1 y_2 \dots y_n. \end{cases}$$

Таким чином, завдяки процесу вирішення рівнянь і виявлення значення формул алгебри скінченних предикатів ідентифікуються близькі за змістом сутності.

Визначено семантико-граматичні характеристики знаків сутностей у вигляді парадигматичних таблиць (табл. 1 – для української мови, табл. 2 – для англійської мови), що зв'язують характеристики слів з можливою їх роллю в субстантивних (два зв'язані іменники), атрибутивних (головним словом виступає іменник, а залежним – прикметник) і дієслівних (дієслово – головний колокат, іменник – залежний колокат) словосполученнях (колокаціях) рефератів. В даних таблицях x визначає головне, а y – залежне слова колокацій, де тип колокації $x_1 y_1$ – субстантивний, $x_2 y_2$ – атрибутивний, $x_3 y_3$ – дієслівний.

Таблиця 1 – Парадигматична таблиця семантико-граматичних характеристик знаків сутностей української мови

c	a	x_1	y_1	y_2	x_2	x_3		y_3
		N	N	A	N	V		N
						Ref	NonRef	
Ag	Nom	q^1		q^{12}	q^{18}	q^{24}	q^{25}	
Att	Gen	q^2	q^7	q^{13}	q^{19}			q^{26}
Pac	Acc	q^3	q^8	q^{14}	q^{20}			q^{27}
Adr	Dat	q^4	q^9	q^{15}	q^{21}			q^{28}
Ins	In	q^5	q^{10}	q^{16}	q^{22}			q^{29}
M	Prt	q^6	q^{11}	q^{17}	q^{23}			q^{30}

Для опису семантичних і граматичних відношень в українській мові введено предметні змінні: a_1 визначає частину мови (N – іменник, A – прикметник, V – дієслово); a_2 визначає відмінок іменників N і прикметників A (Nom – називний, Gen – родовий, Acc – знахідний, Dat – давальний, In – орудний, Prt – місцевий); a_3 визначає зворотність дієслова V (Ref – зворотне, $NonRef$ – незворотне дієслово); c визначає можливі семантичні ролі слів-іменників N (Ag – агенс, Att – атрибут, Pac – пацієнс, Adr – адресат, Ins – інструмент, M – місце).

Таблиця 2 – Парадигматична таблиця семантико-граматичних характеристик знаків сутностей англійської мови

$a \backslash c$	x_1		y_1	y_2		x_2		x_3		y_3
	N		N	A		N		V		N
	$Sub / SubOf$	$ObjOf$	Obj	Att	Pr	$Sub / SubOf$	$ObjOf / Obj$	Tr	$Intr$	$ObjOf / Obj$
Ag	q^1			q^{17}	q^{18}	q^{19}		q^{30}	q^{31}	
Att	q^2	q^7	q^{12}			q^{20}	q^{25}			q^{32}
Pac	q^3	q^8	q^{13}			q^{21}	q^{26}			q^{33}
Adr	q^4	q^9	q^{14}			q^{22}	q^{27}			q^{34}
Ins	q^5	q^{10}	q^{15}			q^{23}	q^{28}			q^{35}
M	q^6	q^{11}	q^{16}			q^{24}	q^{29}			q^{36}

Для формального опису семантико-граматичних відношень англійської мови введено подібні предметні змінні: a_1 визначає частину мови (N – іменник, A – прикметник, V – дієслово); a_2 визначає синтаксичну роль іменника N (Sub – суб’єкт, $SubOf$ – суб’єкт з прийменником «of», Obj – об’єкт, $ObjOf$ – об’єкт з прийменником «of»), a_3 визначає синтаксичну роль прикметника A (Att – атрибутивний, Pr – предикативний), a_4 визначає вид дієслова V (Tr – транзитивне дієслово, $Intr$ – нетранзитивне дієслово).

У парадигматичних табл. 1 та табл. 2 за допомогою формальних номерів $q = \overline{1,30}$ (в українській мові) і $q = \overline{1,36}$ (в англійській мові) позначаються можливі узгоджені значення граматичних і семантичних характеристик слів (потенційних колокатів):

$$\begin{aligned}
 q^1 &= a_1^N a_2^{Nom} c^{Ag}, q^2 = a_1^N a_2^{Gen} c^{Att}, q^3 = a_1^N a_2^{Acc} c^{Pac}, q^4 = a_1^N a_2^{Dat} c^{Adr}, \\
 q^5 &= a_1^N a_2^{In} c^{Ins}, q^6 = a_1^N a_2^{Prt} c^M, q^7 = a_1^N a_2^{Gen} c^{Att}, q^8 = a_1^N a_2^{Acc} c^{Pac}, \\
 q^9 &= a_1^N a_2^{Dat} c^{Adr}, q^{10} = a_1^N a_2^{In} c^{Ins}, q^{11} = a_1^N a_2^{Prt} c^M, q^{12} = a_1^A a_2^{Nom}, q^{13} = a_1^A a_2^{Gen}, \\
 q^{14} &= a_1^A a_2^{Acc}, q^{15} = a_1^A a_2^{Dat}, q^{16} = a_1^A a_2^{In}, q^{17} = a_1^A a_2^{Prt}, q^{18} = a_1^N a_2^{Nom} c^{Ag}, \\
 q^{19} &= a_1^N a_2^{Gen} c^{Att}, q^{20} = a_1^N a_2^{Acc} c^{Pac}, q^{21} = a_1^N a_2^{Dat} c^{Adr}, q^{22} = a_1^N a_2^{In} c^{Ins}, \\
 q^{23} &= a_1^N a_2^{Prt} c^M, q^{24} = a_1^V a_3^{Ref}, q^{25} = a_1^V a_3^{NonRef}, q^{26} = a_1^N a_2^{Gen} c^{Att}, \\
 q^{27} &= a_1^N a_2^{Acc} c^{Pac}, q^{28} = a_1^N a_2^{Dat} c^{Adr}, q^{29} = a_1^N a_2^{In} c^{Ins}, q^{30} = a_1^N a_2^{Prt} c^M;
 \end{aligned} \tag{1}$$

$$\begin{aligned}
 q^1 &= a_1^N a_2^{SubOf} c^{Ag}, q^2 = a_1^N a_2^{SubOf} c^{Att}, q^3 = a_1^N a_2^{SubOf} c^{Pac}, q^4 = a_1^N a_2^{SubOf} c^{Adr}, \\
 q^5 &= a_1^N a_2^{SubOf} c^{Ins}, q^6 = a_1^N a_2^{SubOf} c^M, q^7 = a_1^N a_2^{ObjOf} c^{Att}, q^8 = a_1^N a_2^{ObjOf} c^{Pac}, \\
 q^9 &= a_1^N a_2^{ObjOf} c^{Adr}, q^{10} = a_1^N a_2^{ObjOf} c^{Ins}, q^{11} = a_1^N a_2^{ObjOf} c^M, q^{12} = a_1^N a_2^{Obj} c^{Att}, \\
 q^{13} &= a_1^N a_2^{Obj} c^{Pac}, q^{14} = a_1^N a_2^{Obj} c^{Adr}, q^{15} = a_1^N a_2^{Obj} c^{Ins}, q^{16} = a_1^N a_2^{Obj} c^M, \\
 q^{17} &= a_1^A a_2^{Att}, q^{18} = a_1^A a_2^{Pr}, q^{19} = a_1^N a_2^{SubOf} c^{Ag}, q^{20} = a_1^N a_2^{SubOf} c^{Att}, \\
 q^{21} &= a_1^N a_2^{SubOf} c^{Pac}, q^{22} = a_1^N a_2^{SubOf} c^{Adr}, q^{23} = a_1^N a_2^{SubOf} c^{Ins}, q^{24} = a_1^N a_2^{SubOf} c^M, \\
 q^{25} &= a_1^N a_2^{ObjOf} c^{Att}, q^{26} = a_1^N a_2^{ObjOf} c^{Pac}, q^{27} = a_1^N a_2^{ObjOf} c^{Adr}, q^{28} = a_1^N a_2^{ObjOf} c^{Ins}, \\
 q^{29} &= a_1^N a_2^{ObjOf} c^M, q^{30} = a_1^V a_4^{Tr}, q^{31} = a_1^V a_4^{Intr}, q^{32} = a_1^N a_2^{ObjOf} c^{Att}, \\
 q^{33} &= a_1^N a_2^{ObjOf} c^{Pac}, q^{34} = a_1^N a_2^{ObjOf} c^{Adr}, q^{35} = a_1^N a_2^{ObjOf} c^{Ins}, q^{36} = a_1^N a_2^{ObjOf} c^M.
 \end{aligned} \tag{2}$$

Систему логіко-алгебраїчних рівнянь (1), (2) представлено у спрощеному нормалізованому вигляді (3 – для української мови, 4 – для англійської мови) з урахуванням залежності змінної r від предметних змінних, що виражають семантико-граматичні характеристики колокатів a_1, a_2, a_3, c :

$$\begin{aligned} r^1 &= a_1^N a_2^{Nom} c^{Ag}, r^2 = a_1^N a_2^{Gen} c^{Att}, r^3 = a_1^N a_2^{Acc} c^{Pac}, r^4 = a_1^N a_2^{Dat} c^{Adr}, \\ r^5 &= a_1^N a_2^{In} c^{Ins}, r^6 = a_1^N a_2^{Prt} c^M, r^7 = a_1^A a_2^{Nom}, r^8 = a_1^A a_2^{Gen}, r^9 = a_1^A a_2^{Acc}, \\ r^{10} &= a_1^A a_2^{Dat}, r^{11} = a_1^A a_2^{In}, r^{12} = a_1^A a_2^{Prt}, r^{13} = a_1^V a_3^{Ref}, r^{14} = a_1^V a_3^{NonRef}; \end{aligned} \quad (3)$$

$$\begin{aligned} r^1 &= a_1^N a_2^{SubOf} c^{Ag}, r^2 = a_1^N a_2^{SubOf} c^{Att}, r^3 = a_1^N a_2^{SubOf} c^{Pac}, r^4 = a_1^N a_2^{SubOf} c^{Adr}, \\ r^5 &= a_1^N a_2^{SubOf} c^{Ins}, r^6 = a_1^N a_2^{SubOf} c^M, r^7 = a_1^N a_2^{ObjOf} c^{Att}, r^8 = a_1^N a_2^{ObjOf} c^{Pac}, \\ r^9 &= a_1^N a_2^{ObjOf} c^{Adr}, r^{10} = a_1^N a_2^{ObjOf} c^{Ins}, r^{11} = a_1^N a_2^{ObjOf} c^M, r^{12} = a_1^A a_2^{Att}, \\ r^{13} &= a_1^A a_2^{Pr}, r^{14} = a_1^V a_4^{Tr}, r^{15} = a_1^V a_4^{Intr}. \end{aligned} \quad (4)$$

Побудовано парадигматичні табл. 3 (для української мови), табл. 4 (для англійської мови), в яких жирним шрифтом виділено семантико-граматичні характеристики слів, що стоять поруч і утворюють колокації.

Таблиця 3 – Парадигматична таблиця семантико-граматичних характеристик сутностей, що утворюють колокації української мови

c	a	x_1	y_1	y_2	x_2	x_3		y_3
		N	N	A	N	V		N
						Ref	NonRef	
Ag	Nom	r_x^1		r_y^7	r_x^1	r_x^{13}	r_x^{14}	
Att	Gen	r_x^2	r_y^2	r_y^8	r_x^2			r_y^2
Pac	Acc	r_x^3	r_y^3	r_y^9	r_x^3			r_y^3
Adr	Dat	r_x^4	r_y^4	r_y^{10}	r_x^4			r_y^4
Ins	In	r_x^5	r_y^5	r_y^{11}	r_x^5			r_y^5
M	Prt	r_x^6	r_y^6	r_y^{12}	r_x^6			r_y^6

Таблиця 4 – Парадигматична таблиця семантико-граматичних характеристик сутностей, що утворюють колокації англійської мови

a \ c	x_1		y_1	y_2		x_2		x_3		y_3
	N		N	A		N		V		N
	Sub / SubOf	ObjOf	Obj	Att	Pr	Sub / SubOf	ObjOf / Obj	Tr	Intr	ObjOf / Obj
Ag	r_x^1			r_y^{12}	r_y^{13}	r_x^1		r_y^{14}	r_y^{15}	
Att	r_x^2	r_x^7	r_y^7			r_x^2	r_x^7			r_x^7
Pac	r_x^3	r_x^8	r_y^8			r_x^3	r_x^8			r_x^8

Продовження таблиці 4

$a \backslash c$	x_1		y_1	y_2		x_2		x_3		y_3
	N		N	A		N		V		N
	$Sub / SubOf$	$ObjOf$	Obj	Att	Pr	$Sub / SubOf$	$ObjOf / Obj$	Tr	$Intr$	$ObjOf / Obj$
Adr	r_x^4	r_x^9	r_y^9			r_x^4	r_x^9			r_x^9
Ins	r_x^5	r_x^{10}	r_y^{10}			r_x^5	r_x^{10}			r_x^{10}
M	r_x^6	r_x^{11}	r_y^{11}			r_x^6	r_x^{11}			r_x^{11}

Введено предикати ідентифікації колокацій, що характеризуються системою бінарних відношень (5 – для укр. мови, 6 – для англ. мови):

$$\begin{aligned}
 P(x, y) &= P_1(r_x, r_y) \wedge P_2(r_y, r_x) \wedge P_3(r_x, r_y) = \\
 &= (x^{NNomAg} \vee x^{NGenAtt} \vee x^{NAccPac} \vee x^{NDatAdr} \vee x^{NInIns} \vee x^{NPrtM}) y^{NGen} \vee \\
 &\vee y^{ANom} x^{NNomAg} \vee y^{AGen} x^{NGenAtt} \vee y^{AAcc} x^{NAccPac} \vee y^{ADat} x^{NDatAdr} \vee y^{AIn} x^{NInIns} \vee \\
 &\vee y^{APrt} x^{NPrtM} \vee x^{VNonRef} y^{NAccPac}, \quad (5)
 \end{aligned}$$

$$\begin{aligned}
 P(x, y) &= P_1(r_x, r_y) \wedge P_2(r_y, r_x) \wedge P_3(r_x, r_y) = \\
 &= y^{NObjAtt} x^{NSubAg} \vee (x^{NSubOfAg} \vee x^{NObjOfAtt} \vee x^{NObjOfPac} \vee x^{NObjOfAdr} \vee x^{NObjOfIns} \vee \\
 &\vee x^{NObjOfM}) y^{NObjAtt} \vee y^{AAtt} (x^{NSubAg} \vee x^{NSubOfAg} \vee x^{NObjAtt} \vee x^{NObjPac} \vee x^{NObjAdr} \vee \\
 &\vee x^{NObjIns} \vee x^{NObjM}) \vee x^{NSubAg} y^{APr} \vee x^{VTr} y^{NObjPac}. \quad (6)
 \end{aligned}$$

Предикат $P(x_i, y_i) = 1$ в тому випадку, якщо комбінація семантико-граматичної інформації двох словоформ, які стоять поруч, утворює колокацію, в іншому випадку $P(x_i, y_i) = 0$ (табл. 5 – для укр. мови, 6 – для англ. мови).

Таблиця 5 – Значення предиката $P(x_i, y_i)$ ідентифікації сутностей, що утворюють колокації української мови

$x_i \backslash y_i$	$ANom_a$	$AGen_a$	$AAcc_a$	$ADat_a$	AIn_a	$APrt_a$	$NGen_c$	$NAcc_c$	$NDat_c$	NIn_c	$NPrt_c$
$_aNNom_cAg$	1	0	0	0	0	0	1	0	0	0	0
$_aNGen_cAtt$	0	1	0	0	0	0	1	0	0	0	0
$_aNAcc_cPac$	0	0	1	0	0	0	1	0	0	0	0
$_aNDat_cAdr$	0	0	0	1	0	0	1	0	0	0	0
$_aNIn_cIns$	0	0	0	0	1	0	1	0	0	0	0
$_aNPrt_cM$	0	0	0	0	0	1	1	0	0	0	0
$_aVRef$	0	0	0	0	0	0	0	0	0	0	0
$_aVNonRef$	0	0	0	0	0	0	0	1	0	0	0

Таблиця 6 – Значення предиката $P(x_i, y_i)$ ідентифікації сутностей, що утворюють колокації англійської мови

$x_i \backslash y_i$	$NObjOf/Obj_c Att$	$NObjOf/Obj_c Pac$	$NObjOf/Obj_c Adr$	$NObjOf/Obj_c Ins$	$NObjOf/Obj_c M$	$AAtt$	APr
$_a NSub/SubOf_c Ag$	1	0	0	0	0	1	1
$_a NObjOf/Obj_c Att$	1	0	0	0	0	1	0
$_a NObjOf/Obj_c Pac$	1	0	0	0	0	1	0
$_a NObjOf/Obj_c Adr$	1	0	0	0	0	1	0
$_a NObjOf/Obj_c Ins$	1	0	0	0	0	1	0
$_a NObjOf/Obj_c M$	1	0	0	0	0	1	0
$_a VTr$	0	1	0	0	0	0	0
$_a VIntr$	0	0	0	0	0	0	0

Введено предикат семантичної еквівалентності двослівних колокацій

$$P(x_1, y_1) * P(x_2, y_2) = \gamma_1(x_1, y_1, x_2, y_2) \wedge P(x_1, y_1) \wedge P(x_2, y_2),$$

де знак $*$ позначає операцію визначення смислової близькості; предикат $\gamma_1(x_1, y_1, x_2, y_2)$ виключає колокації, між якими не може бути встановлена смислова еквівалентність.

Предикат $\gamma_1(x_1, y_1, x_2, y_2) = x_1^{VNonRef} y_1^{NAccPac} x_2^{VNonRef} y_2^{NAccPac}$ встановлює семантичну близькість дієслівних колокацій української мови. Наприклад, *зберігати дані* $\approx$ *тримати показники*; *визначати відомості* $\approx$ *встановлювати дані*.

Предикат $\gamma_2(x_1, y_1, x_2, y_2) = x_1^{NNomAg} y_1^{NGenAtt} x_2^{NNomAg} y_2^{NGenAtt}$ визначає семантичну близькість субстантивних колокацій української мови, таких як *набір приладдя* $\approx$ *комплект устаткування*; *процес утворення* $\approx$ *хід формування*.

Предикат $\gamma_3(x_1, y_1, x_2, y_2) = y_1^{ANom} x_1^{NNomAg} x_2^{NNomAg} y_2^{NGenAtt} \vee y_1^{ANom} x_1^{NNomAg} y_2^{ANom} x_2^{NNomAg}$ встановлює семантичну близькість між атрибутивними або атрибутивною та субстантивною колокаціями української мови, наприклад, *грошовий переказ* $\approx$ *відправлення коштів*; *інформаційний потік* $\approx$ *кількість інформації*.

Предикат $\gamma_4(x_1, y_1, x_2, y_2) = x_1^{VTt} y_1^{NObjPac} x_2^{VTt} y_2^{NObjPac}$ встановлює семантичну близькість між дієслівними колокаціями англійської мови: *provide aid* $\approx$ *give support*.

Предикат $\gamma_5(x_1, y_1, x_2, y_2) = x_1^{NSubOfAg} y_1^{NObjAtt} y_2^{NObjAtt} x_2^{NSubAg} \vee x_1^{NSubOfAg} y_1^{NObjAtt} x_2^{NSubOfAg} y_2^{NObjAtt} \vee y_1^{NObjAtt} x_1^{NSubAg} y_2^{NObjAtt} x_2^{NSubAg}$ визначає се-

мантичну близькість субстантивних колокацій англійської мови: *the application of information ≈ the data usage; a content provider ≈ a maintenance supplier*.

Предикат $\gamma_6(x_1, y_1, x_2, y_2) = y_1^{AAtt} x_1^{NSubAg} x_2^{NSubAg} y_2^{APr} \vee$
 $\vee y_1^{AAtt} x_1^{NSubAg} y_2^{AAtt} x_2^{NSubAg} \vee x_1^{NSubAg} y_2^{APr} x_2^{NSubAg} y_2^{APr}$ встановлює семантичну близькість між атрибутивними колокаціями англійської мови: *a wireless device ≈ a cordless machine; a standard is created ≈ the criterion is formed*.

В результаті для ідентифікації семантично зв'язних фрагментів слабо-формалізованої реферативної інформації розроблено логіко-лінгвістичну модель.

У **третьому розділі** представлено методологію процедури визначення єдиних інформаційних просторів наукової взаємодії авторів, що включає встановлення імпліцитних смислових відношень між сутностями знань і компарацію смислових фрагментів рефератів наукометричних систем.

Для ідентифікації знань у наукометричних системах удосконалено метод формалізації смислових відношень сутностей. Для визначення класів еквівалентності та толерантності використовується міра смислової близькості, яка виражається через відношення теоретико-множинного перетину та об'єднання множин термінів дефініцій глосаріїв

$$f(t', t'') = \frac{2 |d_1 \cap d_2|}{|d_1| + |d_2|}, \quad (7)$$

де $d_1 \cap d_2$ – спільні терміни дефініцій глосарію, $|d_1| + |d_2|$ – всі терміни дефініцій d_1 та d_2 .

Для більш точного визначення семантичної близькості використовується декілька глосаріїв. У такому випадку можливі різні дефініції одних і тих самих термінів, при цьому відстань між двома термінами матиме вигляд

$$f(t', t'') = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f(d_{1i}, d_{2j}), \quad (8)$$

де n_1, n_2 – кількість дефініцій першого та другого термінів глосаріїв, d_{1i}, d_{2j} – i -а та j -а дефініції першого та другого терміну, відповідно.

Для визначення міри семантичної близькості (7), (8), при якій смислова кореляція термінів формалізується відношеннями еквівалентності й толерантності, досліджено вибірку з тисячі термінів глосаріїв з інформаційних технологій. В результаті експерименту автоматично виявлено 5000 семантично близьких термінів і визначено коефіцієнт семантичної близькості між ними (табл. 7).

Експертний аналіз показав, що семантичними еквівалентами можуть вважатися ті слова, міра семантичної близькості яких знаходиться в діапазоні від 0,36 до 1. Терміни, зв'язані відношенням толерантності, потрапляють в один клас при значенні коефіцієнта семантичної близькості 0,28–0,35. Саме до цієї категорії входять відношення гіперонімів, гіпонімії, холонімії, міронімії. Семантичною зв'язністю термінів, у яких міра семантичної близькості нижче 0,28, можна знехтувати.

Таблиця 7 – Близькі за смыслом терміни

Коефіцієнт семантичної близькості	Кількість зв'язних фрагментів (%)	Приклад семантично близьких фрагментів рефератів
0,81–0,9	5 (0,1 %)	конкурентне програмування ~ паралельне програмування; transfer speed ~ baud rate
0,51–0,8	95 (1,9 %)	істинна похибка ~ абсолютна похибка; visual attribute ~ graphical characteristic
0,36–0,5	300 (6 %)	математичне забезпечення ~ програмне забезпечення; deep computing ~ profound calculating
0,28–0,35	750 (15 %)	розмірність ~ вимір; Boolean operation ~ logical procedure
0,21–0,27	2250 (45 %)	реєнтерабельна програма ~ повторне використання; system ~ scheme
0,05–0,2	1600 (32 %)	середовище ~ моніторинг; metadata ~ repository

Для виділення категорій відношень в просторі термінів класу толерантності ідентифіковано явні категорії семантичної кореляції за допомогою шаблонів лексичних послідовностей, які представляються регулярними виразами.

Для визначення єдиних інформаційних просторів наукової взаємодії авторів застосовано метод компараторної ідентифікації (рис. 2). При використанні даного методу визначено дві множини: множина рефератів і множина знаків сутностей, виражених лексичними одиницями різних рівнів.

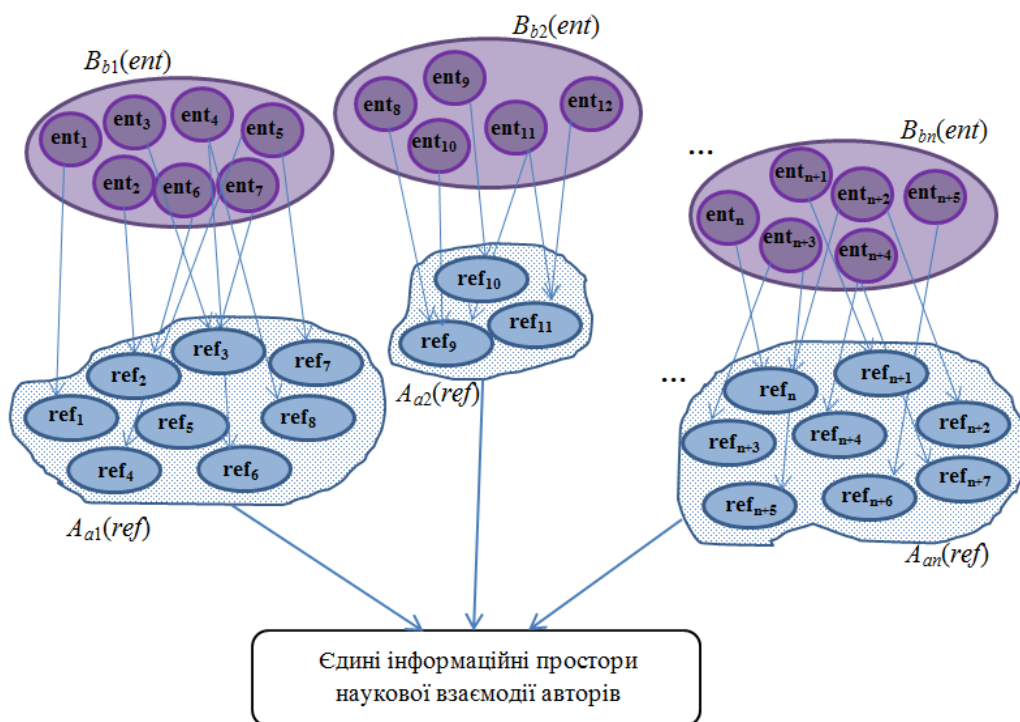


Рисунок 2 – Схема компараторної ідентифікації смислових фрагментів рефератів

При ідентифікації всіх можливих пар з множин *ent* і *ref* компаратор формує контекстно-знаковий предикат, який задає відношення між рефератами і фрагментами інформації, що відображають зміст цих рефератів. Якщо контекстно-знаковий предикат дорівнює 1, то виявлені смислові фрагменти рефератів утворюють інформаційний простір близьких за змістом рефератів. В іншому випадку предикат приймає значення 0. Таким чином, компаратор перевіряє збіг смислу ключових фрагментів та виділяє єдині інформаційні простори.

Четвертий розділ присвячено удосконаленню інформаційної технології ідентифікації знань у наукометричних системах, яка відрізняється етапами визначення імпліцитних зв'язків між рефератами наукометричних систем, що дозволяє динамічно виявляти спільні фронти наукових досліджень (рис. 3).

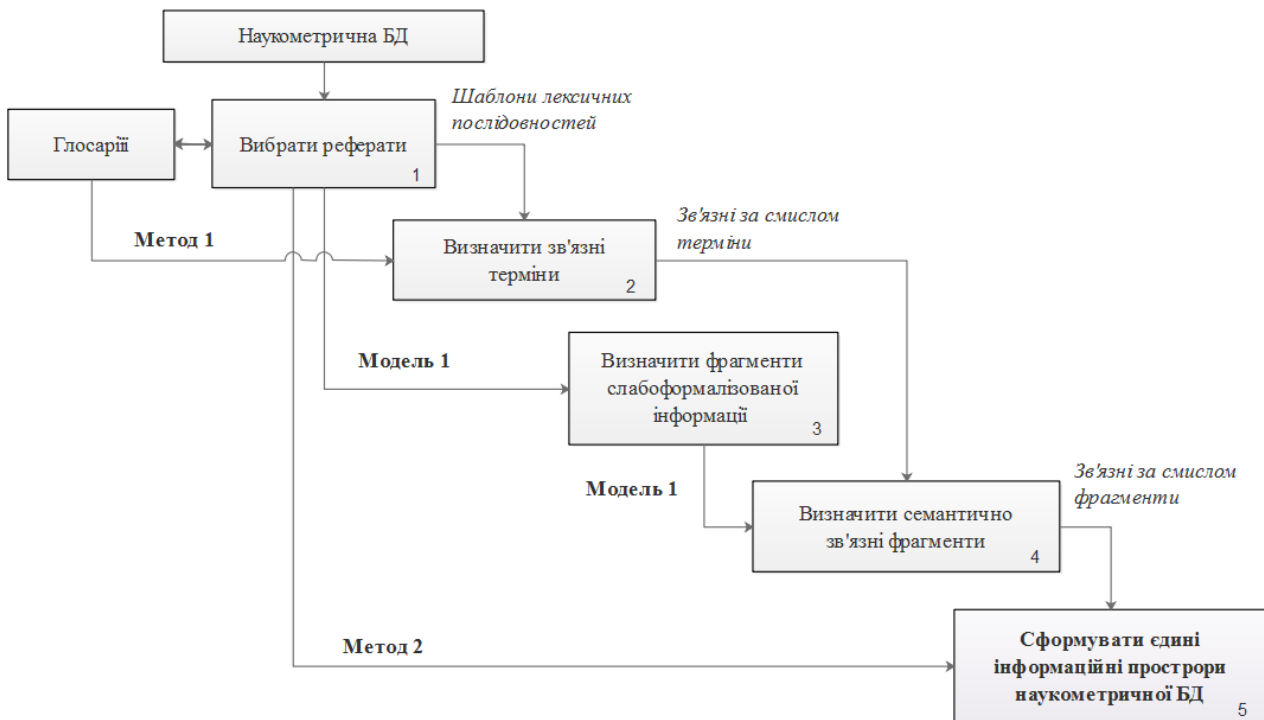


Рисунок 3 – Основні етапи інформаційної технології ідентифікації знань у наукометричних системах

Запропонована інформаційна технологія ідентифікації знань у наукометричних системах дозволяє формувати єдині інформаційні простори наукової взаємодії авторів, включаючи наступні етапи:

- вибір рефератів наукометричної бази, ґрунтуючись на використанні глосарію;
- використання методу формалізації смислових відношень еквівалентності та толерантності для визначення зв'язних термінів;
- виділення семантико-граматичних характеристик слів, що стоять поруч у реченні, та ідентифікація фрагментів слабоформалізованої текстової інформації, представлених субстантивними, атрибутивними і дієслівними колокаціями;
- визначення семантичної еквівалентності виділених фрагментів;
- застосування методу компарації для визначення єдиних інформаційних просторів наукової взаємодії авторів семантично зв'язних рефератів.

У якості об'єктивно вимірюваних показників, що характеризують ефективність розробленої технології, використано коефіцієнти повноти, точності та F_1 -міра, які затверджені міждержавним стандартом з інформації, бібліотечної та видавничої справи

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall},$$

де *precision* – відношення коректно виділених семантично зв'язних сутностей до всіх виділених системою семантично зв'язних сутностей; *recall* – відношення коректно виділених семантично зв'язних сутностей до всіх наявних семантично зв'язних сутностей.

Для розробленого методу формалізації смислових відношень толерантності та еквівалентності сутностей F_1 -міра для англійських рефератів дорівнює 0,9584, а для українських рефератів – 0,9525.

Результати експериментального оцінювання ефективності розробленої моделі представлені в табл. 8.

Таблиця 8 – Оцінка ефективності розробленої моделі ідентифікації семантично зв'язних фрагментів слабоформалізованої інформації

	Розроблена модель		Модель Hua Wu, Ming Zhou	Модель M. Pasca, P. Dienes	Модель R. Barzilay, K. McKeown
	для укр. мови	для англ. мови			
<i>precision</i>	0,7459	0,8793	0,7397	0,4574	0,716
<i>recall</i>	0,8185	0,9398	0,6597	0,7745	0,69
<i>F₁-measure</i>	0,7805	0,9085	0,6974	0,5751	0,7028

В результаті проведеного порівняльного аналізу виявлено, що показники *precision*, *recall* та *F₁-measure* розробленої моделі ідентифікації семантично зв'язних фрагментів слабоформалізованої інформації перевищують аналогічні показники ефективності існуючих моделей.

ВИСНОВКИ

У дисертаційній роботі вирішена науково-практична задача підвищення ефективності процесу ідентифікації знань у наукометричних системах за рахунок побудови моделей і методів інтелектуального аналізу слабоформалізованих даних.

У процесі виконання дисертаційної роботи отримані такі результати.

1. Проаналізовано сучасний стан проблеми виявлення знань у наукометричних системах, систематизовано існуючі засоби інтелектуального аналізу слабоформалізованої інформації, що дозволило сформулювати основні вимоги до розробки інформаційної технології ідентифікації знань.

2. Обґрунтовано використання інструментарію алгебри скінченних предикатів у інформаційно-логічних моделях інтелектуального аналізу, що дозволило побудувати модель ідентифікації знань у наукометричних системах.

3. Побудовано логіко-лінгвістичну модель визначення семантично зв'язних фрагментів реферативної інформації, яка відрізняється застосуванням алгебро-предикатних операцій для формалізації семантико-граматичних характеристик сутностей та дозволяє вдосконалити інформаційну технологію ідентифікації знань у слабоформалізованих даних.

4. Удосконалено метод формалізації смислових відношень сутностей в реферативних текстових потоках, який базується на використанні міри смислової близькості і відрізняється застосуванням інтелектуального аналізу при виявленні класів еквівалентності та толерантності, що дозволяє визначити неявно виражені відношення близькості й таксономії.

5. Отримав подальший розвиток метод компараторної ідентифікації смислових фрагментів рефератів, що дозволяє виділити єдині інформаційні простори наукової взаємодії авторів за рахунок моделювання функцій інтелекту з розуміння та класифікації смислу.

6. Удосконалено інформаційну технологію ідентифікації знань у наукометричних системах, яка відрізняється етапами визначення імпліцитних зв'язків між рефератами наукометричних систем, що дозволяє динамічно виявляти спільні фронти наукових досліджень.

7. Результати дисертаційної роботи впроваджені у системах обробки анотацій та рефератів у Науковій бібліотеці Харківського національного університету радіоелектроніки, Науково-технічній бібліотеці НТУ «ХП», Науковій бібліотеці Національної юридичної академії України ім. Ярослава Мудрого. Теоретичні результати дисертації впроваджені у навчальний процес на кафедрі інтелектуальних комп'ютерних систем НТУ «ХП».

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Петрасова С.В. Метод автоматической экстракции парадигматических отношений между понятиями толкового словаря / С.В. Петрасова, З.А. Кочуева, Н.Ф. Хайрова // Вісник Національного технічного університету «Харківський політехнічний інститут». – Харків : НТУ «ХП», 2013. – № 62 (1035). – С. 118–124.

Здобувачем запропоновано метод автоматичної екстракції смислових відношень між сутностями предметної області.

2. Петрасова С.В. Использование метода компараторной идентификации для динамического наполнения тезауруса оперативно-розыскной деятельности / Н.Ф. Хайрова, Д.Ю. Узлов, С.В. Петрасова // Східно-Європейський журнал передових технологій. – Харків : ПП «Технологічний Центр», 2014. – № 3/2 (69). – С. 4–8.

Здобувачем обґрунтовано використання методу компарації для ідентифікації смислової класифікації лінгвістичних одиниць.

3. Петрасова С.В. Метод автоматической идентификации семантических корреляций терминов глоссария / Н.Ф. Хайрова, С.В. Петрасова, С.В. Ленков // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К. : ВІКНУ, 2014. – № 46. – С. 222–228.

Здобувач запропонувала метод ідентифікації смислових кореляцій термінів глосарія.

4. Петрасова С.В. Логико-лингвистическая модель идентификации семантически эквивалентных коллокаций / С.В. Петрасова, Н.Ф. Хайрова // Вісник Національного технічного університету «Харківський політехнічний інститут». – Харків : НТУ «ХПІ», 2015. – № 58 (1167). – С. 14–17.

Здобувачем запропонована модель автоматичної ідентифікації семантично еквівалентних колокацій української мови.

5. Petrasova S. The logical and linguistic model for automatic extraction of collocation similarity / N. Khairova, S. Petrasova, Ajit Pratap Singh Gautam // Econtechmod : An International Quarterly Journal on Economics in Technology, New Technologies and Modelling Processes. – Lublin, Rzeszow : Polish Academy of Science, 2015. – Vol. 4. – № 4. – P. 43–48.

Здобувачем запропонована модель автоматичної ідентифікації семантично еквівалентних колокацій української та англійської мов.

6. Петрасова С.В. Математическая модель идентификации семантических отношений толерантности и эквивалентности / С.В. Петрасова, Н.Ф. Хайрова // International Journal of Information Theories & Application (ITHEA). – Bulgaria, 2015. – Vol 2. – № 4. – P. 337–344.

Здобувачем наведено метод експериментального визначення ефективності математичної моделі ідентифікації смислових кореляцій.

7. Petrasova S. The Logical-Linguistic Model of Fact Extraction from English Texts / N. Khairova, S. Petrasova, Ajit Pratap Singh Gautam // Communications in Computer and Information Science. – Switzerland : Springer International Publishing, 2016. – № 639. – P. 625–635.

Здобувачем наведено основні засоби формалізації опису семантико-граматичних характеристик слів в англійських реченнях.

8. Петрасова С.В. Автоматизована побудова семантичної мережі на основі тлумачного словника [Електронний ресурс] / С.В. Петрасова, Н.Ф. Хайрова // Горизонти прикладної лінгвістики та лінгвістичних технологій «MegaLing» : тези Міжнар. наук. конф., 20–21 лист. 2013 р., м. Київ. – Режим доступу : [http://megaling.ulif.org. ua/tezi-2013-rik/](http://megaling.ulif.org.ua/tezi-2013-rik/).

Здобувачем розроблено та імплементовано алгоритм виявлення відношень семантичної еквівалентності й таксономії.

9. Петрасова С.В. Лингвистические проблемы использования открытой формы тестовых заданий / С.В. Петрасова, Н.Ф. Хайрова // Экспертные оценки элементов учебного процесса : матер. XV межвуз. научно-практ. конф., 29 нояб. 2013 г. – Харьков : НУА, 2013. – С. 57–60.

Здобувачем запропоновано метод автоматизованої побудови фрагмента семантичної мережі на базі використання знань тлумачного словника.

10. Петрасова С.В. Формализация семантических отношений концептов толкового словаря / С.В. Петрасова // Актуальні проблеми прикладної лінгвістики очима наукової молоді : матер. III Регіон. наук.-практ. конф., 3 груд. 2013 р. – Харків : НАУ «ХАІ», 2013. – С. 40–41.

11. Петрасова С.В. Семантико-когнитивный метод выявления знаний глоссария / С.В. Петрасова // Радиоэлектроника и молодежь в XXI веке : матер. 18-го Междунар. молод. форума, 14–16 апр. 2014 г. – Харьков : ХНУРЭ, 2014. – С. 10–11.

12. Петрасова С.В. Використання методу автоматичної екстракції відношень семантичної близькості для розробки баз знань / С.В. Петрасова // Інтелектуальні системи та прикладна лінгвістика : матер. III Всеукр. наук.-практ. конф., 17 квіт. 2014 р. – Харків : НТУ «ХПІ», 2014. – С. 19–20.

13. Petrasova S. Evaluating Effectiveness of Linguistic Technologies of Knowledge Identification in Text Collections / N. Khairova, G. Shepelyov, S. Petrasova // Transactions on Business and Engineering Intelligent Applications : proc. of 7th International Conference on Intelligent Information and Engineering Systems, 16–20th Sept. 2014. – Rzeszow, 2014. – Vol. 29. – P. 71–75.

Здобувачем запропоновано використання методу тестових колекцій для оцінки розміру експериментальної вибірки текстів.

14. Петрасова С.В. E-learning: формирование единого образовательного пространства на базе контента пользователей социальных сетей / С.В. Петрасова, Н.Ф. Хайрова // Экспертные оценки элементов учебного процесса : матер. XVI межвуз. научно-практ. конф., 26 нояб. 2014 г. – Харьков : НУА, 2014. – С. 61–63.

Здобувачем обґрунтовано використання близьких за змістом сутностей в інформаційних повідомленнях для виявлення загального контенту.

15. Петрасова С.В. E-communities как источник данных в неформальном информационном пространстве / С.В. Петрасова // Інтелектуальні системи та прикладна лінгвістика : матер. IV Всеукр. наук.-практ. конф., 16 квіт. 2015 р. – Харків : НТУ «ХПІ», 2015. – С. 21–22.

16. Петрасова С.В. Автоматическая идентификация семантических корреляций в текстах / С.В. Петрасова, Н.Ф. Хайрова // Інформаційні технології : наука, техніка, технологія, освіта, здоров'я : тези доп. XXIII Міжнар. наук.-практ. конф., 20–22 трав. 2015 р. – Харків : НТУ «ХПІ», 2015. – Ч. 1. – С. 24.

Здобувачем запропоновано метод ідентифікації семантико-синтаксичних кореляцій.

17. Petrasova S. Automatic Identification of Collocation Similarity / S. Petrasova, N. Khairova // Computer Science & Information Technologies (CSIT'2015) : proc. of 10th International Scientific and Technical Conf., 14–17 Sept. 2015. – Lviv, 2015. – P. 136–138.

Здобувачем запропоновано аналіз засобів побудови моделі ідентифікації семантично еквівалентних колокацій української мови.

18. Петрасова С.В. Формалізація інформаційно-лінгвістичних сутностей та відношень засобами штучного інтелекту / С.В. Петрасова // Інтелектуальні

системи та прикладна лінгвістика : матер. V Всеукр. наук.-практ. конф., 14 квіт. 2016 р. – Харків : НТУ «ХПІ», 2016. – С. 48–49.

19. Петрасова С.В. Інтелектуалізація сучасного інформаційного простору / С.В. Петрасова, Н.Ф. Хайрова // Інформаційні технології : наука, техніка, технологія, освіта, здоров'я : тези доповідей XXIV Міжнар. наук.-практ. конф., 18–20 трав. 2016 р. – Харків : НТУ «ХПІ», 2016. – Ч. 4. – С. 302.

Здобувач запропонувала технологію ідентифікації знань в інформаційному просторі.

АНОТАЦІЇ

Петрасова Світлана Валентинівна. Інформаційна технологія ідентифікації знань у наукометричних системах на основі інтелектуального аналізу слабоформалізованих даних. – 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, 2017.

Мета дисертаційного дослідження – підвищення ефективності процесу ідентифікації знань у наукометричних системах за рахунок побудови моделей і методів інтелектуального аналізу слабоформалізованих даних. Основні результати: уперше розроблено логіко-лінгвістичну модель визначення семантично зв'язних фрагментів слабоформалізованої реферативної інформації, яка заснована на використанні алгебро-предикатних операцій, що дозволяє ефективно ідентифікувати знання у реферативній інформації. Удосконалено метод формалізації смислових відношень сутностей, який базується на використанні міри смислової близькості та відрізняється застосуванням інтелектуального аналізу при виявленні класів еквівалентності та толерантності, що дозволяє визначити неявно виражені відношення близькості й відношення таксономії. Отримав подальший розвиток метод компараторної ідентифікації, який використано для класифікації смислових фрагментів рефератів у наукометричних системах що дозволяє виділити єдині інформаційні простори наукової взаємодії авторів за рахунок моделювання функцій інтелекту з розуміння та класифікації смислу. Удосконалено інформаційну технологію ідентифікації знань у наукометричних системах, яка дозволяє за рахунок визначення імпліцитних зв'язків між рефератами наукометричних систем динамічно виявляти спільні фронти наукових досліджень. Результати дослідження знайшли практичне застосування у системах обробки анотацій та рефератів. Використання розроблених у роботі моделей і методів дозволило підвищити ефективність процесу ідентифікації знань у слабоформалізованій реферативній інформації за рахунок підвищення значень коефіцієнтів повноти й точності видачі близької за змістом інформації.

Ключові слова: інформаційна технологія, ідентифікація знань, наукометрична система, логіко-лінгвістична модель, реферативна інформація, смислові відношення.

Petrasova S.V. Information technology of knowledge identification in scientometric systems based on intelligent analysis of weakly formalized data. – Manuscript.

Thesis for a candidate degree in technical sciences, speciality 05.13.06 – Information Technologies. – National Technical University «Kharkiv Polytechnic Institute». – Kharkiv, 2017.

The objective of the thesis is to increase the effectiveness of knowledge identification in scientometric systems by designing the models and methods of intelligent analysis of weakly formalized data.

The main results are as follows. The current state of the knowledge identification problem in scientometric systems has been analysed. Existing methods for the intelligent analysis of weakly formalized data have been systematized. Thus the basic requirements for designing the information technology of knowledge identification have been formulated.

Using the finite predicate algebra in the information and logical models of knowledge identification in Ukrainian and English abstract data of scientometric systems has been proved.

The logical-linguistic model of semantically connected fragments identification in weakly formalized abstract information has been developed. The model is based on the use of algebraic-predicate operations that allows effectively extracting knowledge from abstract information.

The method for the formalization of semantic relations between entities has been improved. The method is based on the use of the semantic similarity measure and intelligent analysis for the identification of equivalence and tolerance classes that allows defining implicit relations of similarity and relations of taxonomy.

The method for comparator identification has got the further development. This method is used to classify abstract fragments in scientometric systems that allows determining common information spaces of scientific interaction by modelling intelligence functions of understanding and classification of sense.

The information technology of knowledge identification in scientometric systems has been improved. The technology allows identifying common research fronts by defining dynamically implicit connections between abstracts of scientometric systems.

The research results have been implemented in the systems of summaries and abstracts processing. Using the developed information technology improves the effectiveness of knowledge identification in weakly formalized data by increasing the average values of the precision and recall measures of semantic similarity of text information.

The practical results can be used in information retrieval, expert, and information-analytical general-purpose systems for the formation of electronic catalogues of semantically connected texts in scientometric, library, and abstract systems.

Keywords: information technology, knowledge identification, scientometric system, logical-linguistic model, abstract information, semantic relations.

Петрасова Светлана Валентиновна. Информационная технология идентификации знаний в наукометрических системах на основе интеллектуального анализа слабоформализованных данных. – 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, 2017.

Цель исследования – повышение эффективности процесса идентификации знаний в наукометрических системах за счет построения моделей и методов интеллектуального анализа слабоформализованных данных. Основные результаты: впервые разработана логико-лингвистическая модель идентификации семантически связанных фрагментов слабоформализованной реферативной информации, основанная на использовании алгебро-предикатных операций, что позволяет эффективно выявлять знания в реферативной информации. Усовершенствован метод формализации смысловых отношений сущностей, основанный на использовании меры смысловой близости и отличающийся применением интеллектуального анализа при выявлении классов эквивалентности и толерантности, что позволяет определить неявно выраженные отношения близости и отношения таксономии. Получил дальнейшее развитие метод компараторной идентификации, который используется для классификации смысловых фрагментов рефератов в наукометрических системах, что позволяет выделить единые информационные пространства научного взаимодействия авторов за счет моделирования функций интеллекта по пониманию и классификации смысла. Усовершенствована информационная технология идентификации знаний в наукометрических системах, что позволяет за счет определения имплицитных связей между рефератами наукометрических систем динамично выявлять общие фронты научных исследований. Результаты исследования нашли практическое применение в системах обработки аннотаций и рефератов. Использование разработанных моделей и методов позволяет повысить эффективность процесса идентификации знаний в слабоформализованной реферативной информации за счет повышения средних значений коэффициентов полноты и точности выдачи близкой по смыслу информации.

Ключевые слова: информационная технология, идентификация знаний, наукометрическая система, логико-лингвистическая модель, реферативная информация, смысловые отношения.



Підписано до друку 23.11.2016 р. Формат 60х90/16.

Гарнітура Times New Roman. Папір офсетний.

Друк – різнограф. Ум. друк. аркушів 0,9

Наклад 100 прим. Зам. № 822242

Надруковано у ТОВ «ПЛАНЕТА-ПРІНТ».

61002, м. Харків, вул. Багалія, 16

ЄДРПОУ 31235131 від 19.12.2000 р.