НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

Каніщева Ольга Валеріївна

УДК 519.7:007.52

ІНФОРМАЦІЙНО-ЛОГІЧНІ МОДЕЛІ І МЕТОДИ ІДЕНТИФІКАЦІЇ ЗНАНЬ В АВТОМАТИЗОВАНИХ ІНФОРМАЦІЙНИХ БІБЛІОТЕЧНИХ СИСТЕМАХ

Спеціальність 05.13.06 – інформаційні технології

Автореферат

дисертації на здобуття наукового ступеня

кандидата технічних наук

Харків – 2010

Дисертацією є рукопис.

Робота виконана на кафедрі інтелектуальних комп'ютерних систем Національного технічного університету "Харківський політехнічний інститут" Міністерства освіти і науки України, м. Харків

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
Замаруєва Ірина Вікторівна,
Військовий інститут Київського національного
університету імені Тараса Шевченка, м. Київ,
професор кафедри
інформаційно-психологічного протиборства

доктор технічних наук, професор
Шабанов-Кушнаренко Сергій Юрійович,
Харківський національний університет
радіоелектроніки, м. Харків,
провідний науковий співробітник кафедри
програмного забезпечення ЕОМ

Захист відбудеться « 25 » березня 2010 р. о 13.00 годині на засіданні спеціалізованої вченої ради Д 64.050.07 в Національному технічному університеті «Харківський політехнічний інститут» за адресою: 61002, м. Харків, вул. Фрунзе, 21.

З дисертацією можна ознайомитись у бібліотеці Національного технічного університету «Харківський політехнічний інститут» за адресою: 61002, м. Харків, вул. Фрунзе, 21.

Автореферат розісланий «_____» _____ 2010 р.

Вчений секретар
спеціалізованої вченої ради В.П. Северин

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Аналіз існуючих завдань, які містять різні аспекти обробки даних, накопичених в електронному вигляді, на перший план виносить проблему обробки текстової інформації у всіх сферах діяльності людини. Однією з важливих сфер її діяльності є система бібліотечно-інформаційного обслуговування. Сучасна бібліотека – це знання-орієнтована інформаційна система, яка оперує з природно-мовними об’єктами, що зберігаються у текстовому вигляді, але недостатня кількість перевірених на практиці теоретичних розробок та ефективних програмно-апаратних систем комп'ютерної обробки знань складає певну проблему.

Теорія і практика створення та використання систем, які засновані на знаннях, інтенсивно розвиваються за напрямком computer science. Навіть на світовому рівні досі недостатньо отримано суттєвих результатів у широких прикладних областях знань, хоча є вдалі рішення для вузькоспеціалізованих застосувань. Зокрема, останнім часом використовуються рішення, відомі під загальною назвою Data mining. Вони дозволяють вилучати з сирих неструктурованих даних за допомогою математичних методів (моделювання, прогнозування, кластеризації, класифікації тощо) раніше невідомі нетривіальні практично корисні й доступні для інтерпретацій нові знання. З появою Інтернету й систем електронного документообігу все більша кількість інформації стала зберігатися в текстовому вигляді. Це призвело до появи рішень для обробки текстової інформації – Text mining, які здійснюють за допомогою лінгвістичних методів наступні дії: тематичний пошук у текстах, класифікація та тематичне індексування документів, відповідь на запити, пошук за ключовими словами, виявлення об'єктів і зв'язків між ними, анотування, реферування і т. ін.

У межах окресленої проблеми важливими є наукові задачі розробки моделей, методів, алгоритмів та програм, які здійснюють моделювання процесів інтелектуальної обробки даних повнотекстових документів з метою визначення їх основних характеристик для побудови інформаційного, математичного, лінгвістичного і програмного забезпечення автоматизованих інформаційних бібліотечних систем (АІБС). У вирішенні задачі ідентифікації знань істотний внесок внесли вчені В.М. Глушков, А.К. Жолковський, Ю.М. Марчук, М. Мінський, О.В. Палагін, Д.О. Поспєлов, Р.Ш. Рубашкін, Ч.Дж. Філлмор, Н. Хомський, Р. Шенк та інші.

Теоретичні аспекти моделювання природної мови та автоматизованої обробки тексту знайшли відображення у працях: Ю.П. Шабанова-Кушнаренка, Ю.Д. Апресяна, М.Ф. Бондаренка, Я.Л. Шрайберга, В.Є. Ярушека, В.О. Широкова, Н.В. Шаронової, І.В. Замаруєвої та ін.

Усе вищезазначене обумовлює актуальність розвитку моделей та методів інтелектуальної обробки даних і застосування цих методів для автоматизованої обробки мовної інформації в бібліотечних системах, що складає напрямок дисертаційної роботи.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана на кафедрі інтелектуальних комп'ютерних систем НТУ "ХПІ" у межах держбюджетної теми МОН України "Розробка математичних моделей та методів розв'язання задач інтелектуальної обробки інформації" (ДР № 0108U003926), у якій здобувач був виконавцем.

Мета і задачі дослідження. Метою дисертаційної роботи є підвищення ефективності та якості інформаційної технології інтелектуальної обробки даних у автоматизованих інформаційних бібліотечних системах на основі інформаційно-логічних моделей та методів ідентифікації знань. Відповідно до зазначеної мети поставлено такі задачі:

1) проаналізувати наукові досягнення в області автоматизації бібліотечних процесів і сформулювати основні вимоги до розробки лінгвістичного забезпечення бібліотечних систем;

2) розробити математичні та лінгвістичні засоби для розв'язання задач обробки текстових документів на основі інтелектуального аналізу даних;

3) розробити модель знання-орієнтованого синтаксичного аналізу у задачах анотування та реферування повнотекстових документів;

4) удосконалити модель процесу систематизації та рубрикації повнотекстових документів в бібліотечних системах;

5) розробити прикладну інформаційну технологію для ідентифікації знань у автоматизованих інформаційних бібліотечних системах;

6) впровадити результати дисертаційної роботи у практику створення інформаційних бібліотечних систем.

Об'єктом дослідження є процеси ідентифікації знань в автоматизованих інформаційних бібліотечних системах.

Предметом дослідження є інформаційно-логічні моделі, що застосовуються у методах ідентифікації знань з використанням інтелектуальних компонент.

Методи дослідження засновані на комплексному використанні теорії інтелекту, алгебри скінченних предикатів та предикатних операцій, методу компараторної ідентифікації, Text Mining для розробки інформаційно-логічних моделей та методів ідентифікації знань. Алгебра предикатів та предикатних операцій використовується для формалізації знань, опису природно-мовних відношень та моделювання синтаксичного розбору у задачах анотування та реферування повнотекстових документів. Метод компараторної ідентифікації використовується для опису інтелектуальних функцій користувача бібліотеки.

Наукова новизна отриманих результатів визначається наступним:

– *Уперше* розроблено метод моделювання процесу ідентифікації знань в електронних бібліотеках, заснований на інтелектуальному аналізі даних, який відрізняється від існуючих комплексним застосуванням логічно пов'язаних методів компараторної ідентифікації, алгебри предикатів і логічних мереж, що дозволяє підвищити ефективність інтелектуальної обробки даних у інформаційних бібліотечних системах.

– *Одержали подальший розвиток* математичні перетворення у алгебрі предикатних операцій шляхом розробки логічної мережі синтаксичних

моделей, що дозволило створити єдину математичну модель ефективної обробки текстової інформації у документах.

– *Удосконалено* інформаційно-логічні моделі процесів систематизації та рубрикації повнотекстових документів за рахунок застосування словника-тезаурусу предметної області, що дозволяє автоматизувати обробку текстової інформації та підвищити швидкість пошуку документів у бібліотеках.

Практичне значення отриманих результатів. Розроблені у дисертації математичні моделі, методи, алгоритми й програмні системи призначені для створення логічних мереж, які є основою для обчислювальних модулів паралельної дії. Математичні результати роботи можуть бути використані в системах автоматичної обробки природної мови, при розробці різних інформаційно-пошукових, експертних, аналітичних засобів інформаційних систем широкого призначення.

Результати дисертаційного дослідження знайшли практичне застосування в науковій бібліотеці Харківського національного медичного університету (м. Харків), науковій бібліотеці Національної юридичної академії України імені Ярослава Мудрого (м. Харків), науковій бібліотеці Харківського національного університету радіоелектроніки (м. Харків), науковій бібліотеці Національного технічного університету "ХПІ" (м. Харків) у вигляді інформаційно-логічних моделей та програмного комплексу, які були використані при розробці автоматизованої інформаційної бібліотечної системи, а також для тематичної систематизації та рубрикації повнотекстових документів. Теоретичні результати дисертації використовуються в навчальному процесі на кафедрі інтелектуальних комп'ютерних систем НТУ «ХПІ» при викладанні спеціальних дисциплін «Інформаційно-ресурсне забезпечення лінгвістичної діяльності», «Автоматизована обробка природної мови» для спеціальності «Прикладна лінгвістика».

Програмне забезпечення, розроблене у дисертації, використовується при виконанні курсових й дипломних робіт на кафедрі інтелектуальних комп'ютерних систем НТУ «ХПІ».

Особистий внесок здобувача. Усі основні результати дисертаційної роботи, що виносяться на захист, отримані здобувачем особисто, серед них: підхід до використання методів Text Mining та Data Mining для обробки текстової інформації в АІБС; використання алгебри предикатів та предикатних операцій для представлення знань; математична модель синтаксичного аналізу речення у задачах анотування та реферування текстової інформації; алгоритм для автоматизованого індексування повнотекстових документів ключовими словами; удосконалення моделі процесу систематизації та рубрикації повнотекстових документів у бібліотечній діяльності.

Апробація результатів дисертації. Результати дисертаційної роботи доповідались та обговорювались на: Міжнародній конференції «Інформаційні технології в освіті і управлінні» (Нова Каховка, 2006, 2007, 2008), X Міжнародній конференції Української асоціації дистанційної освіти (Харків-Ялта, 2006), Міжнародній конференції «MegaLing'2006 Горизонти прикладної лінгвістики та лінгвістичних технологій» (Крим, Партеніт, 2006), Міжнародній

науково-практичній конференції «Інформаційні технології: наука, техніка, технологія, освіта, здоров'я» (Харків, 2006, 2009), Міжвузівському науково-практичному семінарі «Комбінаторні конфігурації та їх застосування» (Кіровоград, 2006, 2007), IX Міжнародній науково-практичній конференції "Системний аналіз та інформаційні технології» (Київ, 2007), I Міжнародній науково-технічній конференції «Інтелектуальні системи в промисловості і освіті-2007» (Суми, 2007), Міжнародному молодіжному форумі «Радіоелектроніка й молодь в 21 столітті» (Харків, 2008, 2009), III міжнародній конференції «Комп'ютерні науки та інформаційні технології» (Львів, 2008), V Міжнародній науково-практичній конференції «Військова освіта і наука: сьогодення та майбутнє» (Київ, 2009), на семінарах кафедри інтелектуальних комп'ютерних систем НТУ «ХПІ».

Публікації. Основні результати дисертації опубліковані у 19 наукових працях, серед яких 8 статей у фахових наукових виданнях ВАК України.

Структура та обсяг дисертації. Дисертаційна робота складається зі списку використаних скорочень, вступу, п'ятьох розділів, висновків, додатків, списку використаних джерел. Повний обсяг дисертації складає 176 сторінок, з них 32 рисунки по тексту, 1 таблиця по тексту, 2 додатки на 11 сторінках, 150 найменувань використаних літературних джерел на 15 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** шляхом аналізу відомого математичного та лінгвістичного забезпечення сучасних автоматизованих інформаційних бібліотечних систем 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 та їх висвітлення у друкованих працях.

У **першому розділі** на основі аналітичного огляду літературних джерел проведено аналіз існуючого математичного та лінгвістичного забезпечення АІБС; зроблено огляд існуючих задач у бібліотечній діяльності, пов'язаних з представленням знань і обробкою текстової інформації. Проаналізовано основні інформаційно-логічні моделі представлення знань, розглянуто основні методи ідентифікації знань (класифікація, кластерізація, семантичні мережі, вилучення фактів тощо). Використання сучасних методів ідентифікації знань дозволяє ефективніше вилучати знання з бібліотечних електронних сховищ. Обґрунтовано необхідність розвитку математичного апарату моделювання інтелектуальних функцій людини для формалізації обробки повнотекстових документів в АІБС. Сформульовано постановку задач подальших досліджень.

На підставі критичного аналізу існуючих невирішених проблем та задач автоматизації бібліотечної діяльності сформульовано й обґрунтовано необхідність дослідження інформаційно-логічних моделей та методів ідентифікації знань для впровадження їх в автоматизовані інформаційні бібліотечні системи, а саме моделей та методів на етапах синтаксичного

аналізу, індексування документів, рубрикації та класифікації документів (рис. 1).

Рис. 1. Структура семантичної обробки інформаційних ресурсів бібліотеки

Другий розділ дисертації присвячено розробці математичного апарату для моделювання процесів обробки повнотекстових даних як основи побудування інформаційної технології. Визначено математичний інструментарій для ідентифікації знань з природно-мовних текстів на основі алгебри предикатів та предикатних операцій, а саме: показано моделювання знань на основі алгебри предикатів, моделі відношень на базі алгебри предикатних операцій, використання алгебри предикатних операцій для опису природно-мовних відношень у задачах ідентифікації знань, моделювання синтаксичної сполучуваності елементів речення.

Предикат – це деяка функція $P(x_1, x_2, ..., x_m)$, яка задана на предметному просторі та відображає цей простір у множину $B = 0,1$, де 0 и 1 – це булеві значення. Під предметним простором у цьому випадку розуміється декартовий добуток $S = A_1 \times A_2 \times ... \times A_m$ довільних непустих не обов'язково різних множин $A_1, A_2, ..., A_m$, які є підмножинами деякого універсуму предметів. Множини $A_1, A_2, ..., A_m$ формують координатні осі простору S.

Значеннями змінної x_i $(i = \overline{1,m})$ є елементи множини A_i $(x_1 \in A_1, x_2 \in A_2, ..., x_m \in A_m)$. Множини $A_1, A_2, ..., A_m$ є областями задання змінних $x_1, x_2, ..., x_m$. Кожній змінній $x_1, x_2, ..., x_m$ ставиться у відповідність фіксована область завдання A_i. Змінні $x_1, x_2, ..., x_m$ змістовно інтерпретуються як місця простору S, а їхні значення (предмети) – як стан цих місць.

Формально операцію заміни відношення предикатом можливо записати наступним чином:

Символ Q у цьому випадку позначає відношення, а символ P – предикат. Зворотній перехід від предиката до відношення формально можна записати так. Якщо $P(x_1, x_2, ..., x_m) = 1$, то $(x_1, x_2, ..., x_m) \in P$, інакше якщо $P(x_1, x_2, ..., x_m) = 0$, то $(x_1, x_2, ..., x_m) \notin P$. Схему переходу від довільного відношення Q до предикату P і навпаки представлено на рис. 2.

Рис. 2. Схема переходу від довільного відношення до предикату

Таким чином, будь-якому відношенню Q у взаємно однозначну відповідність ставиться предикат, який на мові алгебри предикатів записується у формульному вигляді, використовуючи базисні предикати 0, 1 та предикат упізнавання предмету x_i^a, а також базисні операції кон'юнкції $\wedge$ та

диз'юнкції $\vee$. Предикат упізнавання предмета a, який знаходиться на місці x_i, визначається наступною формулою:

Доведено, що на мові формул алгебри предикатів можна представити будь-які відношення. Відокремлений факт, який представлено набором $(a_1, a_2, ..., a_m)$ станів $a_1, a_2, ..., a_m$ міст $x_1, x_2, ..., x_m$ можна виразити висловленням «$x_1 = a_1$ і $x_2 = a_2$ і... і $x_m = a_m$», якому відповідає формула алгебри предикатів $x_1^{a_1} \wedge x_2^{a_2} \wedge ... \wedge x_m^{a_m}$. Формули такого виду називаються конституентами одиниць предикату. Довільне знання про факт представляється відношенням $\{(a_{11}, a_{21}, ..., a_{m1}), (a_{12}, a_{22}, ..., a_{m2}), ..., (a_{1r}, a_{2r}, ..., a_{mr})\}$, де r – це число наборів у відношенні. Це знання виражається висловленням

Алгебра предикатів описує тільки знання про факт та формує лише те, що називають висловленням, судженням. Алгебра операцій формалізує операції над знанням, які представлені у вигляді відношень на деякому предметному просторі S. Алгебра предикатів описує декларативну складову знань, а алгебра предикатних операцій – процедурну складову знань. Алгебра предикатних операцій служить для реалізації виводу знань, отримання нових знань на основі множини аксіом або первинних знань, які зберігаються у базі знань інтелектуальної інформаційної системи.

Алгебра предикатів та предикатних операцій у теорії інтелекту може бути використана для опису законів поведінки випробовуваного, для опису баз даних та баз знань, а також для опису структури речень і тексту. Показано, що текст – це формула, яка записана мовою лінгвістичної алгебри предикатних операцій.

Якщо записати мовою алгебри предикатних операцій висловлення "Результати проведеного дослідження впроваджено у практику охорони здоров'я", то слова, що входять до складу цього речення, можна розглядати як факти, а речення – як знання про факти. Мовою лінгвістичної алгебри операцій речення запишеться наступним чином

де $X_1, \ldots, X_8$ – це предикати, що відповідають словам розглянутого речення. Мовою алгебри предикатів предикати-слова виражаються у вигляді формул:, , …,. Відношення, які зафіксовані у цьому реченні, записуються у формульному вигляді

(1)

Лінгвістичне рівняння (1) розв'язує програмним шляхом за допомогою відповідного модуля розв'язання лінгвістичних рівнянь, або представляється у вигляді лінгвістичного перемикального ланцюга, який можна реалізувати апаратно, використовуючи розроблені останнім часом мікроелектронною

промисловістю програмувальні логічні інтегральні мікросхеми. Перемикальний ланцюг, що відповідає лінгвістичному рівнянню (1), представлений на рис. 3.

Таким чином, з використанням алгебри предикатів та предикатних операцій створена інтегрована модель представлення знань, яка базується на традиційних моделях та моделях представлення знань природною мовою.

У **третьому розділі** розглядається застосування методів Text Mining для ідентифікації знань в АІБС. Технологія Text Mining – це інструментарій, який дозволяє аналізувати великі обсяги інформації у пошуку тенденцій, шаблонів та взаємозв'язків, які можуть допомогти у прийнятті стратегічних рішень. Крім того, Text Mining – це перспективний вид пошуку, який, на відміну від традиційних підходів, не тільки знаходить списки документів, формально релевантних запитам, але й забезпечує достатньо високий рівень аналізу з метою прийняття ефективного рішення.

Рис. 3. Перемикальний ланцюг для рівняння (1)

У роботі розв'язана задача індексування повнотекстових документів ключовими словами з використанням методів Text Mining. Ця задача безпосередньо пов'язана з питанням пошуку ключових слів та словосполучень у тексті. Індексування текстів є однією з проблем при створенні інформаційно-пошукових систем, які використовують у якості критеріїв пошуку набір ключових слів. На рис. 4 представлені етапи пошуку ключового слова або словосполучення у повнотекстовому документі.

Індексування текстів є однією з проблем при створенні інформаційно-пошукових систем, що використовують як критерії пошуку набори ключових слів та словосполучень. У запропонованому підході спочатку проводиться пошук іменників, знайдені іменники оброблюються за допомогою морфологічного аналізу. Для цього алгоритму використано алгоритмічний морфологічний аналіз методів Text Mining.

Морфологічний аналіз здійснено шляхом виділення у складі аналізованої словоформи певної основи слова та певного закінчення слова. Потім відбувається порівняння інформації про основу та закінчення, у результаті якого формується комплекс морфологічної інформації всієї словоформи. Для зручності аналізу всі закінчення кожної частини мови групуються за словозмінними типами. Наприклад, для морфологічного аналізу науково-технічних текстів можна обійтися шістнадцятьма словозмінними типами іменників. Після морфологічного аналізу знайдені слова приводяться до канонічну форму.

Однак не завжди ключові слова представлені одним словом. Програма індексації повнотекстових документів ключовими словами повинна знаходити не тільки ключові слова, але й словосполучення.

Недоліком багатьох програм є те, що в них відсутній пошук ключових слів конструкції "прикметник + іменник", а також не враховуються випадки заміни іменників на займенники при підрахунку кількості повторень слова в тексті.

Для вирішення першої проблеми запропоновано виділити граматичні ознаки за характерним закінченням прикметників і присвоювати ці ознаки іменникам.

Рис. 4. Етапи пошуку ключового слова в повнотекстовому документі

Для вирішення другої необхідно зберігати всі словоформи займенників у базі даних (БД) словника, через те, що в російській та українській мовах мало займенників, а їх вага в частотній таблиці велика. Доцільно також використати невелику базу зі словами, що не є ключовими за визначенням (сполучники, частки тощо), що дозволить не враховувати при підрахунку вагу цих слів. При програмній реалізації у роботі запропоновано використовувати словник "Розділ знань", який полегшує пошук ключових слів у документі й містить синоніми часто вживаних термінів у предметній області.

Розроблено засоби для моделювання процесів систематизації та рубрикації повнотекстових документів у бібліотечній діяльності. Об'єкти, які оброблюються інформаційними системами, є дискретними, скінченними й детермінованими, що дозволило використовувати при обробці цих об'єктів метод компараторної ідентифікації. На вхід системи подається множина сигналів $x_1, x_2, \ldots, x_n$. Під сигналом розуміємо умовні знаки, які служать для передачі інформації (тексти документів, ключові поняття тощо). Вхідні сигнали беруться зі скінченних множин $X_1, X_2, \ldots, X_n$, при цьому $x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n$. У результаті роботи системи, яка здійснює обробку інформації, на вихід поступає певна множина елементів $y_1, y_2, \ldots, y_n$. Під $y_1, y_2, \ldots, y_n$ вважаємо ключові поняття, дескриптори, рубрики, заголовки та підзаголовки документів тощо, при цьому $y_1 \in Y_1, y_2 \in Y_2, \ldots, y_n \in Y_n$.

У ході роботи системи перевіряється існування деякого відношення Q, яке зв'язує елементи $y_1, y_2, \ldots, y_n$, що з'являються на виході системи після сигналів $x_1, x_2, \ldots, x_n$, які поступають на вхід системи. Таким чином, у процесі роботи компаратор реалізує предикат $q = Q(y_1, y_2, \ldots, y_n)$, який відповідає відношенню Q та характеризує механізм порівняння елементів $y_1, y_2, \ldots, y_n$. Предикат $P(x_1, x_2, \ldots, x_n) = Q(f_1(y_1), f_2(y_2), \ldots, f_n(y_n))$ характеризує роботу системи, що здійснює інтелектуальну обробку документальної інформації, яка на сигнали $x_1, x_2, \ldots, x_n$ реагує відповіддю $q = P(x_1, x_2, \ldots, x_n)$. Моделювання будь-якої із задач аналітико-синтетичної обробки інформації полягає у тому, щоб із властивостей предиката P, який здійснює компараторну ідентифікацію інформаційних об'єктів, вилучити внутрішню структуру сигналів $x_1, x_2, \ldots, x_n$, елементів $y_1, y_2, \ldots, y_n$, вид функцій $f_1, f_2, \ldots, f_n$ і вид предиката Q.

Представлено вдосконалену модель процесу систематизації та рубрикації повнотекстових документів у бібліотечній діяльності з використанням словника-тезауруса та методу компараторної ідентифікації. Змінено схему роботи бібліотечного рубрикатора, що дозволило точніше відносити документи

до певних рубрик, а також збільшити множину ключових слів для пошуку відповідного документа (рис. 5).

Рис. 5. Схема роботи бібліотечного рубрикатора

Ця модель вдосконалює не тільки процес систематизації, а й процес інформаційного пошуку у бібліотечно-пошуковій системі.

Четвертий розділ присвячено розробці інформаційної технології морфолого-синтаксичної обробки у задачах анотування та реферування текстових документів. У розділі використано алгебру скінченних предикатів при моделюванні синтаксичного розбору речення, показано скорочення дерев та розроблено логічну мережу для синтаксичного розбору речення.

Представимо опис та синтаксичний розбір речення "*Методы и алгоритмы Data Mining наиболее эффективны при анализе больших объемов данных*" мовою алгебри предикатів та предикатних операцій. Схема формули цього речення показана на рис. 6.

((Методы1(алгоритмы))2(Data Mining))9((наиболее3(эффективны)8((при4(анализе))7((больших5(объемов))6(данных))). (2)

Рис. 6. Схема формули речення

Замінюючи у виразі (2) усі слова відповідними їм предикатними змінними $X_1 \div X_{10}$, приходимо до формули алгебри предикатних операцій

що виражає синтаксичну структуру розглянутого речення.

Нормалізована формула має наступний вигляд

$$((X_1 1 X_2) 2 X_3) 9 ((X_5 3 X_4 8 ((X_6 4 X_7) 7 ((X_8 5 X_9) 6 (X_{10}))).$$

Таким чином, формула речення остаточно побудована та представляє собою формулу алгебри предикатних операцій.

Розроблено логічну мережу для синтаксичного розбору речення у задачах анотування та реферування. Логічна мережа є графічним представленням результату бінарної кон'юнктивної декомпозиції багатомісного предикату. Спочатку користувач (експерт) вводить правила, які описують ситуацію, описує зв'язки між словами (узгодження, керування, підпорядкування), а мережа може "відповідати на питання" (з точністю до вихідних даних), знаходити відсутні слова у реченні. Вихідні дані надходять у відповідні полюси мережі; результат

вирішення також отримується у полюсах після зупинення її роботи. Логічна мережа працює в інтерактивному режимі, кожна ітерація називається тактом. Зупинка відбувається, коли стан мережі на черговому такті повторюється. Представлено логічну мережу для синтаксичного розбору речення типу "Висока частота поширеності захворювань потребує удосконалення знань у цій галузі" (рис. 7).

Рис. 7. Структура логічної мережі

Предметними змінними цього речення є слова самого речення, тобто

Область зміни всіх предметних змінних для мережі синтаксичного розбору речення формується експертом, наприклад,

Бінарні зв’язки між вузлами логічної мережі описують рівняння:

П’ятий розділ присвячено практичній реалізації отриманих результатів з тематичної рубрикації та систематизації повнотекстових документів.

Проведена оцінка ефективності запропонованих семантичних моделей та методів обробки текстової інформації. Розроблена програма для рубрикації медичних повнотекстових документів реалізована за допомогою методу компараторної ідентифікації. Текстовими даними була колекція україномовних та російськомовних повнотекстових документів: електронні ресурси наукової бібліотеки Харківського національного медичного університету. Приклад інтерфейсу програми представлено на рис. 8.

Рис. 8. Інтерфейс програми

Оцінка ефективності рубрикації документів виконана шляхом обчислення мір якості отриманих результатів класифікації (рубрикації). Запропоновано чотири показника якості інформаційних ресурсів по відношенню до включення у відповідь і релевантних рубриці документів: a – кількість документів, релевантних рубриці у пошуковій видачі; b – кількість документів, що не є релевантними рубриці у пошуковій видачі; c – кількість не виданих у пошукову видачу, але релевантних рубриці документів; d – кількість не виданих у пошукову видачу та не релевантних рубриці документів. Було проаналізовано близько 300 документів з 30 рубрик. Використано зовнішні міри для порівняння автоматичної розбивки з одержаною від експертів "еталонною" розбивкою цих же даних для розробленого програмного забезпечення. Для проаналізованих рубрик отримані наступні усереднені значення показників: $\mathrm{Recall} = 0{,}873$, $\mathrm{Precision} = 0{,}913$, $\mathrm{Error} = 0{,}014$, де Recall – міра повноти, Precision – міра точності, Error – помилка автоматичної рубрикації.

Проведені експерименти показали ефективність запропонованого в роботі методу компараторной ідентифікації для рубрикації повнотекстових

документів. Програмна реалізація методу може бути використана для електронних бібліотек як елемент пошукових систем.

У роботі також показані перспективи використання запропонованих моделей і методів ідентифікації знань: використання алгебри скінченних предикатів для автоматизації процедури відмінювання іменників при мінімальному залученні користувача для ідентифікації морфологічних характеристик нового слова при складанні електронних словників; алгебро-логічні моделі емоційно-насиченої лексики для оцінки емоційної насиченості текстів засобів масової інформації по відношенню до деякого об’єкту.

ВИСНОВКИ

У дисертаційній роботі вирішена актуальна науково-практична задача обґрунтування та розробки перспективної інформаційної технології інтелектуальної обробки текстових даних у автоматизованих інформаційних бібліотечних системах на основі інформаційно-логічних моделей та методів ідентифікації знань.

У процесі виконання дисертаційної роботи отримані наступні результати.

1. Проаналізовано сучасні інформаційні бібліотечні системи та сформульовані основні вимоги до їх математичного та лінгвістичного забезпечення. Виявлено особливості тексту як об’єкта моделювання, обробки та представлення знань, проаналізовані існуючі інформаційно-логічні моделі ідентифікації знань.

2. Для моделювання процесів обробки повнотекстових даних обґрунтовано використання алгебри предикатів та предикатних операцій, проведено аналіз їх властивостей для представлення та ідентифікації знань. Створено інтегровану модель представлення знань, яка базується на традиційних логічних моделях та на моделях природної мови, з використанням алгебри предикатів та предикатних операцій.

3. Розроблено модель знання-орієнтованого синтаксичного аналізу у задачах анотування та реферування повнотекстових документів. Розроблено логічну мережу для синтаксичного аналізу на базі моделей сполучуваності слів, яка спирається на структуру речення та семантику тексту в цілому, що дозволило перейти до єдиної математичної моделі обробки текстової інформації у документах з використанням єдиного математичного апарату.

4. Удосконалено модель процесу систематизації та рубрикації повнотекстових документів у бібліотечних системах з використанням словника-тезауруса та методу компараторної ідентифікації, що дозволило точніше сформувати пошуковий образ документа та якісніше за змістом здійснювати рубрикацію документів та пошук у бібліотечній пошуковій системі.

5. Розроблено прикладну інформаційну технологію для ідентифікації знань у автоматизованих інформаційних бібліотечних системах. Проведено оцінку ефективності запропонованих семантичних моделей та методів обробки текстової інформації для реальних АІБС. Виділено і проаналізовано кількісні та

якісні показники підрахунку ефекту від упровадження моделей, методів та алгоритмів.

6. Результати роботи впроваджено при розробці інформаційного, математичного, алгоритмічного й програмного забезпечення у наукову бібліотеку Харківського національного медичного університету, наукову бібліотеку Національної юридичної академії України імені Ярослава Мудрого (м. Харків), наукову бібліотеку Харківського національного університету радіоелектроніки, наукову бібліотеку НТУ "ХПІ" у вигляді інформаційно-логічних моделей, які були використані при розробці автоматизованої підтримки інформаційної бібліотечної системи, а також для тематичної систематизації та рубрикації повнотекстових документів та використані у навчальному процесі на кафедрі інтелектуальних комп'ютерних систем НТУ "ХПІ" при викладанні спеціальних дисциплін «Інформаційно-ресурсне забезпечення лінгвістичної діяльності», «Автоматизована обробка природної мови».

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Канищева О. В. Использование методов Data Mining и Text Mining для обработки текстовой информации в информационных системах / О. В. Канищева, Сайед Мохаммад Таухид Сиддики, Н. В. Шаронова // Біоніка інтелекту. – Харків : ХНУРЕ, 2005. – № 2(63). – С. 22-26.

Здобувачем обґрунтовано підхід до використання методів Data Mining і Text Mining для обробки текстової інформації в інформаційних системах.

2. Канищева О. В. Методы Data Mining в автоматизированных информационных библиотечных системах / Н. В. Шаронова, О. В. Канищева // Вестник Херсонского национального технического университета. – Херсон : ХНТУ, 2006. – № 1(24). – С. 157-162.

Здобувач запропонував метод компараторної ідентифікації як однин з логічних методів Data Mining для обробки повнотекстової інформації.

3. Канищева О. В. Эффективный анализ текстовой информации с помощью технологий Data Mining / Н. В. Шаронова, О. В. Канищева, Сайед Мохаммад Таухид Сиддики // Вісник Національного технічного університету "ХПІ". – Харків : НТУ "ХПІ", 2006. – № 19. – С. 87-92.

Здобувачем запропоновано підхід до використання методів Text Mining для обробки текстової інформації в інформаційних бібліотечних системах.

4. Канищева О. В. Использование алгебры предикатов и предикатных операций для формализации декларативной и процедурной составляющих знаний / З. А. Алисейко, В. И. Булкин, О. В. Канищева, Н. В. Шаронова // Біоніка інтелекту. – Харків : ХНУРЕ, 2006. – № 1(64). – С. 59-63.

Здобувач розробив математичну модель представлення знань з використанням алгебри предикатів та предикатних операцій.

5. Канищева О. В. Автоматизированное индексирование полнотекстовых документов ключевыми словами / З. А. Алисейко, О. В. Канищева // Вестник

Херсонского национального технического университета. – Херсон : ХНТУ, 2007. – № 4(27). – С. 269-272.

Здобувачем розроблено алгоритм для автоматизованого індексування повнотекстових документів ключовими словами.

6. Канищева О. В. Идентификация информационных объектов в современной библиотеке с использованием алгоритма реферирования / О. В. Канищева, З. А. Кочуева, Н. В. Шаронова // Вестник Херсонского национального технического университета. – Херсон : ХНТУ, 2008. – № 1(30). – С. 126-130.

Здобувач запропонував підхід до ідентифікації інформаційних об'єктів у сучасній бібліотеці, на основі використання алгоритму реферування.

7. Канищева О. В. Моделирование синтаксического анализа в задачах аннотирования и реферирования полнотекстовых документов / Н. В. Борисова, О. В. Канищева // Вісник Національного технічного університету "ХПІ". – Харків : НТУ "ХПІ", 2009. – № 4. – С. 87-96.

Здобувачем використана алгебра предикатів та предикатних операцій для моделювання синтаксичного розбору у задачах анотування та реферування.

8. Каніщева О. В. Алгебра скінченних предикатів як складова інформаційних технологій / С. В. Гончаров, О. В. Канищева // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К. : ВІКНУ, 2009. – № 22. – С. 80-84.

Здобувач використав алгебру скінченних предикатів для однакового представлення знань в електронних бібліотеках у вигляді рівнянь алгебри предикатів.

9. Канищева О. В. Моделирование процедур систематизации и предметизации полнотекстовых документов / З. А. Алисейко, О. В. Канищева, Н. В. Шаронова // Проблемы информационных технологий. – Херсон : ХНТУ, 2007. – № 1. – С. 140-144.

Здобувачем використаний метод компараторної ідентифікація в процедурах систематизації та предметизації повнотекстових документів.

10. Канищева О. В. Проблемы индексирования полнотекстовых документов по ключевым словам / О. В. Канищева, Н. В. Шаронова // Библиотеки учебных заведений. Научно-методический журнал для библиотек учебных заведений системы профессионального образования. – М. : ГУП, 2007. – № 23. – С. 25-32. – Режим доступу до журн. : www.lib-journal.ru.

Здобувач обґрунтовано використання методів статистичного контент-аналізу для індексування повнотекстових документів ключовими словами.

11. Канищева О. В. Применение методов Data Mining и Text Mining в автоматизированных информационных библиотечных системах / Н. В. Шаронова, О. В. Канищева // Международная конференция Украинской ассоциации дистанционного образования : материалы 10-й междунар. конф. Украинской ассоциации дистанционного образования. – Харьков-Ялта : ХНУРЕ, 2006. – С. 129-135.

Здобувачем обґрунтовано підхід до використання методу компараторної ідентифікації як одного з логічних методів Text Mining.

12. Канищева О. В. Обработка текстовой информации с помощью технологий Text Mining и компараторной идентификации / Н. В. Шаронова, Сайед Мохаммад Таухид Сиддики, О. В. Канищева // MegaLing'2006 Горизонти прикладної лінгвістики та лінгвістичних технологій : доповіді міжнар. конф., Україна, Крим, Партеніт, 20-27 вересня 2006 р. / Укр. мовно-інформаційний фонд НАН України, Таврійський національний університет ім. В.І. Вернадського. – Сімферополь : Вид-во "ДиАйПи", 2006. – С. 231-232.

Здобувач запропонував використання методу компараторної ідентифікації для обробки повнотекстових документів.

13. Канищева О. В. Идентификация знаний в электронных библиотеках / О. В. Канищева // Системный анализ и информационные технологии : тез. IX Междунар. науч.-практ. конф. – К. : НТУУ "КПІ", 2007. – С. 114.

14. Канищева О. В. Обработка текстовой информации с помощью технологий Text Mining и компараторной идентификации / Н. В. Шаронова, О. В. Канищева // Комбінаторні конфігурації та їх застосування : II міжвуз. наук.-практ. семінар. – Кіровоград : ДЛАУ, 2006. – С. 56-57.

Здобувачем проведено порівняльний аналіз методу компараторної ідентифікації та методів Text Mining для обробки текстової інформації.

15. Канищева О. В. Моделирование процессов реферирования и аннотирования полнотекстовой информации в библиотеках / Н. В. Шаронова, О. В. Канищева // Інтелектуальні системи в промисловості і освіті: тези доп. І міжнар. наук.-техн. конф. – Суми : СумДУ, 2007. – С. 142-143.

Здобувач запропонував модель анотування та реферування з використанням алгебри скінченних предикатів.

16. Канищева О. В. Основы информационной технологии тематического рубрицирования / О. В. Канищева // Радіоелектроніка і молодь в XXI ст. : тези доп. 12-ого міжнар. молод. форуму. – Харків : ХНУРЕ, 2008. – С. 508.

17. Канищева О. В. Багаторівневий підхід до тонального аналізу повнотекстової інформації / О. В. Канищева // Комп’ютерні науки та інформаційні технології : матеріали III міжнар. конф. CSIT'2008. – Львів : Вежа і Ко, 2008. – С. 134-136.

18. Канищева О. В. Проблема пополнения электронных словарей новыми словами / Н. В. Борисова, О. В. Канищева // Радіоелектроніка і молодь в XXI ст. : тези доп. 13-ого міжнар. молод. форуму. – Харків : ХНУРЕ, 2009. – Ч. 2 – С. 91.

Здобувачем запропоновано використовувати апарат алгебри скінченних предикатів для вилучення морфологічних характеристик нового слова у електронних словниках.

19. Канищева О. В. Использование алгебры предикатных операций для описания естественно-языковых отношений / О. В. Канищева // Інформаційні

технології: наука, техніка, технологія, освіта, здоров'я : матеріали XVII міжнар. наук.-практ. конф. – Харків : НТУ "ХПІ", 2009. – С. 16.

Здобувачем використав алгебру предикатних операцій для опису природно-мовних відношень.

АНОТАЦІЇ

Каніщева О.В. - Інформаційно-логічні моделі і методи ідентифікації знань в автоматизованих інформаційних бібліотечних системах. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Національний технічний університет «Харківський політехнічний інститут», Харків, 2010.

У дисертаційній роботі вирішена науково-практична задача підвищення ефективності та якості інтелектуальної обробки даних у автоматизованих інформаційних бібліотечних системах на основі інформаційно-логічних моделей та методів ідентифікації знань.

У роботі проаналізовано наукові досягнення в області автоматизації бібліотечних процесів і сформульовано основні вимоги до розробки лінгвістичного забезпечення бібліотечних систем. Для моделювання процесів обробки повнотекстових даних обґрунтовано використання алгебри предикатів та предикатних операцій, проведено аналіз їх властивостей для представлення та ідентифікації знань. Створено інтегровану модель представлення знань, яка базується на традиційних логічних моделях представлення знань, а також на моделях представлення знань природною мовою, з використанням алгебри предикатів та предикатних операцій.

Розроблено модель знання-орієнтованого синтаксичного аналізу у задачах анотування та реферування повнотекстових документів. Розроблено логічну мережу для синтаксичного аналізу на базі моделей сполучуваності слів, яка спирається на структуру речення та семантику тексту в цілому, що дозволило перейти до єдиної математичної моделі обробки текстової інформації у документах з використанням єдиного математичного апарату. Удосконалено модель процесу систематизації та рубрикації повнотекстових документів у бібліотечних системах з використанням словника-тезауруса та методу компараторної ідентифікації, що дозволило точніше сформувати пошуковий образ документа та якісніше за змістом здійснювати рубрикацію документів та пошук у бібліотечній пошуковій системі.

Ключові слова: автоматизована переробка інформації, інтелектуальна обробка даних, автоматизовані інформаційні бібліотечні системи, алгебра предикатів, алгебра предикатних операцій, компараторна ідентифікація.

Канищева О. В. Информационно-логические модели и методы идентификации знаний в автоматизированных информационных библиотечных системах. – Рукопись.

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

технический университет «Харьковский политехнический институт», Харьков, 2010.

Диссертация посвящена решению научно-практической задачи повышения эффективности и качества интеллектуальной обработки данных в автоматизированных информационных библиотечных системах на основе информационно-логических моделей и методов идентификации знаний.

В исследовании проанализированы научные достижения в области автоматизации библиотечных процессов и сформулированы основные требования к разработке лингвистического обеспечения библиотечных систем.

Разработаны математические и лингвистические средства для решения задач обработки текстовых документов на основе интеллектуального анализа данных. Обосновано использования алгебры предикатов и предикатных операций для моделирования процессов обработки полнотекстовых данных. Создана интегрированная модель представления знаний, которая базируется на традиционных логических моделях представления знаний, а также на моделях представления знаний естественным языком, с использованием алгебры предикатов и предикатных операций.

Разработана модель знание-ориентированного синтаксического анализа в задачах аннотирования и реферирование полнотекстовых документов. Разработана логическая сеть для синтаксического анализа на базе моделей сочетаемости слов, которая опирается на структуру предложения и семантику текста в целом. Усовершенствована модель процесса систематизации и рубрикации полнотекстовых документов в библиотечных системах с использованием словаря-тезауруса и метода компараторной идентификации. В работе показано использование методов Text Mining и метода компараторной идентификации для решения задачи автоматической индексации ключевыми словами полнотекстовых документов.

Разработана прикладная информационная технология для идентификации знаний в автоматизированных информационных библиотечных системах. Проведена оценка эффективности предложенных семантических моделей и методов обработки текстовой информации для библиотечных систем.

Результаты диссертационной работы внедрены в научную библиотеку Харьковского национального медицинского университета, научную библиотеку Национальной юридической академии Украины имени Ярослава Мудрого (г. Харьков), научную библиотеку Харьковского национального университета радиоэлектроники, научную библиотеку НТУ "ХПІ", также в учебном процессе на кафедре интеллектуальных компьютерных систем НТУ "ХПІ" при преподавании специальных дисциплин "Информационно-ресурсное обеспечение лингвистической деятельности", "Автоматизированная обработка естественного языка".

Ключевые слова: автоматизированная переработка информации, интеллектуальная обработка данных, автоматизированные информационные библиотечные системы, алгебра предикатов, алгебра предикатных операций, компараторная идентификация.

Kanishcheva O. V. **Information and logical models of identification of knowledge in the automated information library systems.** – Manuscript.

Thesis for a Candidate Degree in Technical Sciences, Specialty 05.13.06 – Information technologies. – National Technical University «Kharkiv Polytechnic Institute», Kharkiv, 2010.

Dissertational work is devoted to modelling of processes of processing of the text-through information in the automated information library systems. The dissertation focuses on the task of improving the quality and effectiveness of intellectual data processing in automated information library systems on the basis of information and logical models and data identification methods.

The latest scientific achievements in the sphere of automated library systems have been analysed and basic requirements to library systems lingware have been suggested. It has been proved that algebra of final predicates is to be used for modelling the processing of text-through data, their properties have been analysed to introduce and identify knowledge. The integrated model of knowledge introduction has been developed based on traditional logical models of knowledge introduction and on the models of knowledge introduction using natural languages with the help of algebra of final predicates and predicate operations.

The model of knowledge oriented syntactic analysis in annotation and abstracting of text-through documents has been worked out. The logical network for syntactic analysis based on word combination model based on sentences structure and text semantics on the whole, which allowed to proceed to the mathematical model of text processing in the documents with the single mathematical apparatus. The model of systematization and rubrication of text-through documents in library systems using a thesaurus and the method of comparator identification has been improved which allowed to form the search image of a document more exactly and to rubricate and search documents in library search systems with higher quality.

Key words: automated information processing, intellectual data processing, automated information library systems, algebra of predicates, algebra of predicate operations, comparator identification.