

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

Гончаров Сергій Володимирович

УДК 681.518: 519.767

**МЕТОДИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ФОРМУВАННЯ
ГАЛУЗЕВИХ ЕЛЕКТРОННИХ БІБЛІОТЕЧНИХ РЕСУРСІВ**

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
Єрохін Андрій Леонідович,
Харківський національний університет
внутрішніх справ, м. Харків,
начальник кафедри інформатики

доктор технічних наук, професор
Сенченко Микола Іванович,
Книжкова палата України ім. Івана Федорова,
м. Київ, директор

Захист відбудеться "25" березня 2010 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. Фрунзе, 21.

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

Автореферат розісланий 23.02.2010 р.

Вчений секретар
спеціалізованої вченої ради

Северин В. П.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Інформатизація значно розширила можливості соціальних, зокрема наукових, комунікацій і зумовила принципові зміни в діяльності, структурі та функціях бібліотек, які відіграють велику роль в забезпеченні інформаційних потреб різних груп населення. Автоматизація всіх процесів бібліотечної діяльності повинна забезпечити доступ користувача до електронного каталогу та інших розподілених баз даних бібліотек з максимально адекватною відповідністю отриманої інформації до читацьких запитів. Розробка ефективних автоматизованих засобів ідентифікації змісту текстів, і насамперед – україномовних, пов'язаних із зазначеною предметною галуззю, і таких, що входять до складу традиційних та електронних фондів бібліотек, є актуальною задачею, яка відповідає головним завданням Національної програми інформатизації.

Значний науковий і практичний внесок в розвиток галузі автоматизації доступу до інформації в бібліотеках зроблений Ф. Воройським, Р. Гарбором, Л. Костенком, М. Сенченком, Дж. Солтоном, Л. Філіповою, І. Фоменком, Дж. Хеклі, В. Цурканом, Я. Шрайбергом та іншими.

Оскільки більшість документів у бібліотеках представлено природними мовами, актуальною задачею є обробка інформації за допомогою методів штучному інтелекту, створення комп'ютерних систем, що автоматизують інтелектуальну діяльність людини. Істотний внесок в розв'язання питань моделювання розуміння в інтелектуальних системах, в розвиток теорії і практики моделювання функцій інтелекту і властивих йому процесів, зокрема переробки текстів природної мови (ПМ), внесли Т. Віноград, Н. Леонова, О. Палагін, Д. Мінський, С. Осуга, О. Перевозчикова, Е. Попов, Д. Поспелов, Ю. Саєкі, К. Філлмор, Р. Шенк, Н. Хомський, Ю.П. Шабанов-Кушнарєнко, Н.В. Шаронова та інші науковці.

Розвиток методів автоматизації процесів створення галузевих електронних бібліотечних фондів із застосуванням сучасних методів інтелектуальної обробки текстової інформації передбачає ґрунтовне вивчення структури предметної галузі й особливостей мовної поведінки досвідчених фахівців, розробку необхідного математичного, лінгвістичного та програмного забезпечення відповідних підсистем автоматизованої інформаційно-бібліотечної системи (АІБС). Дослідження обумовлено необхідністю підвищення ефективності використання українських інформаційних ресурсів та забезпечення фахівців в реальний проміжок часу повною і релевантною інформацією. Все вищезазначене обумовлює актуальність вибраного напрямку дисертаційного дослідження.

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

програми Tempus/Tacis "Інтеграція у європейську інформаційну мережу навчальних закладів ветеринарного профілю" (№ CD-JEP-21136-2000) у відділі сучасних інформаційних технологій Харківської державної зооветеринарної академії.

Мета і задачі дослідження. Метою дисертаційної роботи є підвищення ефективності формування галузевих електронних бібліотечних ресурсів на основі моделей, методів та інформаційних технологій інтелектуальної обробки текстів.

Відповідно до зазначеної мети необхідно розв'язати наступні задачі:

1) дослідити логіко-семантичну структуру предметної області формування фондів галузевих бібліотек, визначити особливості їхньої автоматизації та застосування математичних моделей процесу аналітико-синтетичної обробки текстів документів;

2) побудувати модель ідентифікації змістовних ознак інформаційних лінгвістичних об'єктів, що входять до складу галузевих фондів, розробити математичну модель інтелектуальної обробки даних, яка дозволяє перейти до знання-орієнтованих методів формування фондів;

3) розробити засоби автоматичної побудови семантичної мережі спеціалізованого україномовного тезаурусу з ветеринарної медицини, систематизації текстів повнотекстової бази даних і пошуку в множині текстів, релевантних запиту користувача, вираженому обмеженою природною мовою;

4) розробити прикладну інформаційну технологію інтеграції розподілених галузевих ресурсів, реалізувати розроблені методи і моделі у вигляді модулю, інтегрованого в підсистему каталогізації і користувацького доступу існуючої бібліотечної інформаційної системи; розробити проблемно-орієнтовані засоби навігації в галузевих інформаційних масивах;

5) розробити електронний галузевий рубрикатор з ветеринарної медицини, його структуру, правила використання та супроводження.

Об'єктом дослідження є процеси формування електронних бібліотечних ресурсів у системі інформаційного забезпечення галузі.

Предмет дослідження – інтелектуальні моделі та методи автоматизації обробки даних при формуванні і використанні інформаційних ресурсів електронної бібліотеки в галузі ветеринарної медицини.

Методи дослідження засновані на комплексному використанні теорії інтелекту, теорії категорій та логічних мереж, апарату алгебри предикатів для моделювання процесу семантичної обробки бібліотечних документів галузевої тематики. Основи теорії множин та теорії відношень використовувались для переходу до алгебро-логічних моделей. Імітаційне моделювання – для практичної реалізації отриманих результатів.

Наукова новизна отриманих результатів визначається наступним:

Вперше запропоновано метод формування інформаційних бібліотечних фондів, заснований на застосуванні методу компараторної ідентифікації, теорії категорій та апарату логічних мереж, що дозволяє підвищити ефективність формування галузевих бібліотечних ресурсів.

Набули подальшого розвитку моделі інтелектуальної обробки інформаційних об'єктів на основі автоматичної індексації і ідентифікації текстів в ієрархічній структурі класифікаційної системи, що дозволило автоматизувати процес створення та поповнення галузевих термінологічних словників.

Удосконалено інформаційну технологію аналітико-синтетичної обробки текстів, шляхом створення методів автоматичної побудови семантичної мережі та україномовного тезаурусу, що дозволило підвищити ефективність доступу до галузевих бібліотечних фондів.

Теоретична значимість отриманих результатів полягає в розробці наукових засад створення та розвитку галузевої інформаційної системи, що передбачає можливість автоматизованого формування розподілених бібліотечних ресурсів, централізоване формування лінгвістичного препроцесору галузевої термінології та системи навігації в ній.

Практичне значення отриманих результатів. Розроблені у дисертації математичні засоби, моделі, алгоритми й програмні системи дозволяють автоматизувати формування галузевих бібліотечних фондів за рахунок визначення та ефективної обробки текстів галузевої спрямованості. Побудовано електронний галузевий рубрикатор, який характеризується тематичною спрямованістю, виконує довідково-пошукову функцію, а також функцію розподілу напрямків предметної області між ветеринарними установами, з одного боку, та об'єднання документно-інформаційних ресурсів – з іншого. Практичні результати дисертаційних досліджень впроваджені для покращення автоматизації аналітико-синтетичної обробки підфону спеціалізованих текстів у науково-методичному відділі наукової бібліотеки Харківського національного медичного університету та для поліпшення якості формування галузевих ветеринарних фондів і підвищення релевантності й повноти обслуговування читачів наукового інформаційно-бібліотечного центру Харківської державної зооветеринарної академії.

Теоретичні результати дисертації використовуються в навчальному процесі на кафедрі інтелектуальних комп'ютерних систем НТУ "ХПІ" при викладанні спеціальних дисциплін "Інформаційно-ресурсне забезпечення лінгвістичної діяльності", "Комп'ютерні системи перекладу і документування інформації" для спеціальності "Прикладна лінгвістика".

Особистий внесок здобувача. Всі результати дослідження отримані автором самостійно. В роботах, що були виконані у співавторстві, дисертантом особисто:

- розроблено формальний опис структури текстів ветеринарної медицини та підхід до використання методів Data Mining і Text Mining для обробки текстової інформації в галузевих фондах;
- запропоновано математична модель аналізу предметної галузі на основі виділення сукупності пов'язаних дисциплін у системах автоматизованого формування бібліотечного фонду;

- обґрунтовано використання методу компараторної ідентифікації для моделювання інтелектуальної обробки текстів галузевої належності і зроблено відповідну доробку математичного апарату;
- розроблена і програмно реалізована модель підсистеми автоматичної індексації, систематизації та предметизації текстів в АІБС для подальшого реферування;
- розроблено структуру та здійснено програмну реалізацію рубрикатора з ветеринарної медицини.

Апробація результатів дисертації. Основні результати дисертаційної роботи доповідалися та обговорювалися на: Міжнародній конференції "Інформаційні технології в освіті і управлінні" (Нова Каховка, 2008, 2009); Науково-методичній конференції "Сучасна електронна бібліотека як складова сучасної системи освіти" (Харків, 2004); Підсумковій міжнародній конференції "TEMPUS 2001 Реформи ветеринарної освіти у Харкові, Україна" (м. Харків, 2004); XV Міжнародній науково-практичній конференції "Інформаційні технології: наука, техніка, технологія, освіта, здоров'я" (м. Харків, 2007); 13 Міжнародному молодіжному форумі "Радіоелектроніка й молодь в 21 столітті" (м. Харків, 2009), Першій міжвузівській науковій конференції "Інтелектуальні системи і прикладна лінгвістика" (Харків, 2009); V Міжнародній науково-практичній конференції "Військова освіта і наука: сьогодення та майбутнє" (Київ, 2009), семінарах кафедри інтелектуальних комп'ютерних систем НТУ "ХПІ".

Публікації. Основні результати дисертації опубліковані в 8 наукових працях, серед яких 6 статей у фахових наукових виданнях ВАК України.

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, додатків, списку використаних джерел. Повний обсяг дисертації становить 191 сторінку, з них 21 рисунок по тексту, 151 найменування використаних літературних джерел на 14 сторінках, 4 додатки на 22 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність теми роботи з точки зору перспектив розвитку та удосконалення систем автоматизованого формування галузевих бібліотечних фондів. Визначено мету роботи та сформульовані задачі дослідження. Охарактеризовано наукову новизну та практичне значення отриманих результатів, наведено інформацію про практичне використання результатів роботи, їх апробацію та публікації.

Перший розділ присвячений аналізу досягнень в області автоматизації бібліотечної діяльності, огляду літературних джерел у галузі автоматизації функцій спеціальних і наукових галузевих бібліотек, досліджень з використання інформаційних технологій у бібліотеках на сучасному етапі, аналізу основних проблем створення галузевих фондів бібліотек, зокрема у галузі ветеринарної медицини та інформаційних технологій формування електронних ресурсів. На підставі критичного аналізу існуючих невирішених

проблем та задач автоматизованого формування електронних фондів сформульовані й обґрунтовані цільові задачі наукових досліджень дисертації.

Бібліотека представляє собою достатньо складну систему, в якій виділено типові підсистеми і процеси. Формування фонду бібліотеки та керування його формуванням ґрунтується на повній інформації про його поточний склад і оцінці інтегрального показнику якості. Широке введення автоматизації доцільно на тих ділянках бібліотечних технологічних процесів, які складаються з послідовності типових взаємопов'язаних процесів: бібліотечна обробка (інвентарний облік, систематизація, складення бібліографічного опису), облік читачів, облік руху документів (рис. 1).

Рис.1. Інформаційно-логічна структура формування фонду бібліотечної системи

У роботі розглянуто зібрання електронних документів в галузі ветеринарної медицини, серед них як документи, які розташовані в Інтернеті, так і спеціальні інформаційні сховища інформації, такі як, наприклад, БД по дисертаціям з ветеринарії. Значну роль в інформаційному забезпеченні наукових досліджень також відіграють централізовані цифрові зібрання матеріалів наукових журналів за різними напрямками. Електронні бібліотеки із розвинутими пошуковими системами, в тому числі – спеціальні і тематичні, є одним із обов'язкових засобів забезпечення дистанційної освіти. У якості галузевих електронних ресурсів у роботі визначена сукупність електронних документів, видань, продуктів або їх комбінацій, які відповідають певній галузі знань та забезпечують повноту відображення та наочність цієї галузі. Галузеві електронні ресурси формуються з використанням інформаційних технологій – цілеспрямовано організованої сукупності інформаційних процесів із використанням засобів обчислювальної техніки.

Показано, що найважливішими задачами формування бібліотечних фондів, які потребують впровадження сучасних інформаційних технологій, є задачі *предметизації* і *класифікації*. Сутність класифікації – у встановленні тотожності об'єктів за певною ознакою або їх сукупністю, яка є основою класифікації. Систематизація документів в галузі бібліотечно-бібліографічної діяльності є реалізацією формального алгоритму класифікації за тією або іншою системою, заходом, який застосовують для групування документів або метаінформаційних записів в каталогах і довідниках. Класифікаційні системи використовують один з двох принципів структурної організації: ієрархічний і фасетний, або їх комбінації. Головними вимогами до об'єктивної системи класифікації є дотримання єдиної основи, вичерпність, взаємовиключення (неперетинання) класів і послідовність. Результатом систематизації є присвоєння документу класифікаційного індексу, що визначає його місце в класифікаційному просторі. В більш складних випадках документ може бути

співставлений із декількома розділами УБК, що відбивається в утворенні складного класифікаційного індексу. Якщо методика систематизації за певною УБК дозволяє складання комплексних індексів (як, наприклад, УДК), то вона набуває рис фасетної класифікації.

Методики спеціальної систематизації документів, пов'язаних із ветеринарною медициною, не є досконалими через складність предметної галузі і необхідність виконання класифікаційного аналізу наукового характеру для кожного документа, що під силу лише фахівцю у галузі. Автоматизація обробки документів предметної області із введенням нових ефективних методів семантичного аналізу вивільняє час досвідчених фахівців із збереженням якості інформаційного забезпечення галузі.

Актуальним аспектом розвитку електронних каталогів є вдосконалення систем автоматичної аналітико-синтетичної обробки текстів електронних документів і пошуку релевантних документів за запитами користувача, що виражені природною мовою, а також підтримка декількох мов. Бібліотечно-бібліографічні класифікації представляють собою індексну підмножину інформаційно-пошукових мов (ІПМ) двох типів: класифікаційні (координатні) мови, що ототожнюють зміст документа із координатами певного класифікаційного простору у вигляді індексів, і дескрипторні, що описують зміст документу лексичними одиницями ІПМ. Використання класифікаційних ІПМ передбачає наявність спеціального лексичного аналізатора, що здійснює, на основі обраної системи правил індексування, розбір індексу на складові і відтворення його координатної позиції в класифікації. Показано, що при здійсненні тематичного пошуку документів в масиві індексів необхідно транслювати запити користувачів в індексну ІПМ, а отримані індекси релевантних документів – в більш зручні для використання і легкі для розуміння речення природної мови. Аналіз предметної області ветеринарної медицини показав високий ступень перетинання тематичних класів, а отже – велике число складених індексів.

У найбільшій мірі повноту відображення та наочність певної галузі знань забезпечують термінологічні словники, які містять визначення основних понять, які складають терміносистему всієї галузі, без якої неможливе існування даної галузі знань. При створенні галузевих електронних ресурсів особлива увага приділяється створенню електронних термінологічних словників. На сучасному етапі розвитку науки в Україні існує реальна потреба у формуванні галузевих електронних ресурсів, для створення яких можна використовувати різні інформаційні технології. У роботі пропонується використання алгебро-логічних методів моделювання мовних відношень мовою рівнянь алгебри скінчених предикатів.

Другий розділ присвячено дослідженню логіко-семантичної структури предметної області формування фондів галузевих бібліотек, визначено особливості їхньої автоматизації та застосування інформаційних технологій у процесі аналітико-синтетичної обробки текстів документів галузевого фонду.

Реалізація ефективного пошуку спеціалізованих текстів вимагає вивчення структури предметної області, знаходження її специфічних семантичних ознак,

вивчення процесу пошуку релевантних джерел фахівцем в галузі та відтворення цього процесу в АІБС. Це можливе лише на основі формальної моделі дії фахівця при аналізі повнотекстової інформації та ідентифікації її змісту. Для її створення пропонується використовувати метод компараторної ідентифікації лінгвістичних об'єктів, який є ефективним засобом опису інтелектуальної діяльності людини. Це дозволяє створити модель розуміння тексту людиною на основі представлення її у вигляді сюр'єктивної функції f , що відображає множину можливих ситуаційних змінних $x_1, x_2, \dots, x_n$ на множину суб'єктивних станів $y_1, y_2, \dots, y_n$ свідомості особи, поведінка якої досліджується. Теорія компараторної ідентифікації дозволила з'ясовувати внутрішню структуру сигналів x_j і y_j ; вигляд функції f_j і вигляд предикату, що описує поведінку особи.

У роботі дослідження та моделювання здійснювалися на прикладі словоформ української мови. Кожна людина є носієм предикатів тих слів, словосполучень та речень, які вона розуміє. Таким чином, смислом речення є предикат, який задає визначений зв'язок $L(X, Y)$ між смислом та відповідним йому фрагментом тексту. Тут $X = (x_1, x_2, \dots, x_m)$ – змінний вектор змісту, m – число його компонент, а $y_1, y_2, \dots, y_n$ – деякий фрагмент тексту, де $y_1, y_2, \dots, y_n$ – букви, які стоять на 1, 2, ..., n місцях фрагмента.

Проаналізована можливість використання предикатних категорій для формалізації знань, представлених за допомогою логічних моделей. У роботі припущено, що АІБС, яка вирішує задачі фактографічного та бібліографічного (документального) пошуку, являє собою логічну модель подання знань, в її основі лежить формальна система M , представлена четвіркою виду: $M = \langle T, P, A, B \rangle$, де T – це множина усіх текстів, які містить система; P – це множина ключових слів певної предметної галузі (це може бути множина рубрик галузевого рубрикатора або галузевий тезаурус тощо); A – це множина текстів певної предметної галузі, які віднесені до однієї рубрики, B – це також підмножина M_{cnc} ($A \subset M_{cnc}$); множина M_{cnc} – множина текстів певної предметної галузі; B – це множина інформаційно-тематичних запитів користувачів, за допомогою яких, з множини A , отримують множину M_{ec} – множину текстів, які відповідають запитам користувачів, до яких знову можна застосовувати запиту з множини B .

Функціонування АІБС можна відтворити у вигляді послідовності наступних відображень. $T \xrightarrow{P} M_{cnc}$ – множина текстів певної предметної

галузі отримується з множини усіх текстів з використанням множини ключових слів певної предметної галузі, або множини рубрик рубрикатора, або галузевого тезаурусу. Множина $M_{cnc} \xrightarrow{P_A} A$ текстів певної предметної галузі, які віднесені до однієї рубрики, отримується з усієї множини текстів предметної галузі за допомогою множини ключових слів певної предметної галузі, які характеризують дану рубрику, або множини рубрик рубрикатора. Множина $T \xrightarrow{P P_A} A$ текстів певної предметної галузі, які віднесені до однієї рубрики, отримується з множини усіх текстів, використовуючи множини ключових слів певної предметної галузі, або множини рубрик рубрикатора, або галузевого

тезаурусу, а на наступному етапі використовуючи ключові слова, які характеризують дану рубрику.

Множину текстів, які відповідають запитам користувачів, можна отримати трьома шляхами. У першому випадку відображення матиме вигляд $A \xrightarrow{B} M_{ec}$. Множину текстів, які відповідають запитам користувачів, можна отримати з множини текстів певної предметної галузі, віднесених до однієї рубрики, за допомогою множини запитів користувачів. У другому випадку відображення матиме вигляд $M_{cnc} \xrightarrow{P_A B} M_{ec}$, тобто необхідну множину можна отримати, використовуючи множину ключових слів певної предметної галузі, які характеризують певну рубрику, або множину рубрик рубрикатора та запити користувачів. І нарешті, при необхідності отримати множину необхідних текстів з множини усіх текстів $T \xrightarrow{(P_A)^B} M_{ec}$ необхідно послідовно скористатися множиною ключових слів певної предметної галузі, або множиною рубрик рубрикатора, або галузевим тезаурусом, множиною ключових слів певної предметної галузі, які характеризують певну рубрику, або множиною рубрик рубрикатора та множиною запитів користувачів.

Далі, застосовуючи запити користувачів з множини B до отриманої множини текстів M_{ec} , які цим запитам відповідають, отримуємо нову множину M_{ec}^1 , до якої, у свою чергу, можна знову застосувати елементи множини B та отримати чергову множину M_{ec}^2 і т.д. Змістова постановка задачі зводиться до того, щоб дати інтерпретацію понять, які формуються теорією категорій, і методів, які нею розроблюються, у термінах алгебри предикатів. Під цим розуміють перехід від абстрактного визначення категорії до конкретних екземплярів об'єктів, морфізмів та ін. у термінах алгебри предикатів, тобто до предикатної інтерпретації категорії або предикатної категорії.

Таким чином, у цьому розділі у якості математичного апарату для моделювання інформаційних процесів обробки текстових даних у документах спеціалізованих фондів обрані метод компараторної ідентифікації, алгебра скінченних предикатів, теорія категорій та логічних мереж. Побудовано формалізовану модель процесу ідентифікації змістовних ознак інформаційних лінгвістичних об'єктів, що входять до складу галузевих фондів. Інформаційну бібліотечну систему представлено за допомогою логічної моделі подання знань. Показано застосування предикатних категорій для формалізації знань, представлених за допомогою логічних моделей. Функціонування логічної моделі подання знань відтворено у вигляді послідовності відображень й категоричних діаграм.

У **третьому розділі** проведено порівняльний аналіз методів інтелектуальної обробки даних. Запропоновано використовувати метод компараторної ідентифікації, як один з логічних методів, близький до методів Text Mining і який є ефективним для моделювання процесів обробки інформаційних об'єктів, що дозволяє моделювати процедури інтелектуальної (семантичної) обробки документів, на відміну від технології Data Mining, яка передбачає аналіз упорядкованої певним чином інформації.

На основі порівняльного аналізу запропоновано використання

формальних засобів компараторної ідентифікації для моделювання аналітико-синтетичної обробки бібліотечних об'єктів: дескрипторно-текстовий предикат, предикат аналітико-синтетичної обробки документа, вивчено їх властивості, потрібні для опису функції розуміння тексту і лексичної одиниці, що виражає певне поняття, з використанням одержаних теоретичних результатів при побудові тематичного рубрикатора з ветеринарної медицини. Показано застосування апарату алгебри скінченних предикатів та методу компараторної ідентифікації на задачах створення та поповнення галузевих термінологічних словників.

Класична задача ідентифікації полягає у тому, що за вхідним x і вихідним у сигналам об'єкта потрібно визначити функцію $y = F(x)$ перетворення сигналу цим об'єктом. Таку ідентифікацію звать прямою, оскільки вона здійснюється при безпосередньому доступі до вихідного сигналу об'єкта. Однак у ряді випадків виникає необхідність у непрямій ідентифікації об'єкта, коли у дослідника немає прямого доступу до вихідного сигналу. Метод компараторної ідентифікації об'єкта дозволяє викладати основні положення теорії інтелекту дедуктивним способом, виходячи виключно з фактів, які можна фізично спостерігати, він добре зарекомендував себе при обробці лінгвістичних об'єктів різних рівнів ПМ.

Об'єкти, які оброблюються інформаційними системами, є дискретними, скінченними й детермінованими, що дозволяє використовувати при обробці цих об'єктів метод компараторної ідентифікації. На вхід системи подається множина сигналів $x_1, x_2, \dots, x_n$. Під сигналом ми розуміємо умовні знаки, які служать для передачі інформації (тексти документів, ключові поняття й т.д.). Вхідні сигнали одержуються зі скінченних множин $X_1, X_2, \dots, X_n$, причому $x_1 \in X_1, x_2 \in X_2, \dots, x_n \in X_n$. У результаті роботи системи, яка здійснює обробку інформації, на вихід поступає визначена множина елементів $y_1, y_2, \dots, y_n$. У нашому випадку під $y_1, y_2, \dots, y_n$ можна розуміти ключові поняття, дескриптори УДК, ББК і т.д. При цьому $y_1 \in Y_1, y_2 \in Y_2, \dots, y_n \in Y_n$. Елементи $y_1, y_2, \dots, y_n$ однозначно залежать від сигналів $x_1, x_2, \dots, x_n$. Це означає, що існують функції $y_1 = f_1(x_1), y_2 = f_2(x_2), \dots, y_n = f_n(x_n)$, які ставлять у відповідність кожному з елементів $x_1 \in X_1, x_2 \in X_2, \dots, x_n \in X_n$ елементи $y_1 \in Y_1, y_2 \in Y_2, \dots, y_n \in Y_n$. При цьому кожному з сигналів $x_i \in X_i$ відповідає цілком визначений елемент $y_i \in Y_i$. Таким чином, кожна з функцій f_i являє собою сюр'єкцію, яка відображає множину X_i на множину $Y_i, i \in \{1, 2, \dots, n\}$. Функції f_i характеризують здібність системи співвідносити інформацію, яка передається умовним знаком – сигналом з елементом, який відображає його зміст й відповідає тій чи іншій класифікації.

У роботі розподіл текстів на підтеми та ключових понять на підрубрики виражено через предикат P , що об'єктивно визначається компаратором. Класу L_a всіх текстів $t \in T$, що відносяться до однієї підтеми, яка містить текст $a \in T$, відповідає предикат $L_a(t)$

$$L_a(t) = E_1(t, a) = \forall r \in R(P(t, r) \sim P(a, r)). \quad (1)$$

Класу Q_b всіх ключових понять $r \in R$, що відносяться до однієї підрубрики з ключовим поняттям $b \in R$, відповідає предикат $Q_b(r)$

$$Q_b(r) = E_2(r, b) = \forall t \in T(P(t, r) \sim P(t, b)). \quad (2)$$

Формули (1) та (2) виражають поділ текстів на підтеми та ключових понять на підрубрики через предикат P , який об'єктивно визначається класифікатором. Розглянуто роботу класифікатора на прикладі предметної галузі ветеринарної медицини. У прикладі припущено, що у повнотекстовій базі даних є 10 текстів документів. Повні бібліографічні описи документів наведені у додатку С. Всі 10 документів віднесені до єдиної рубрики 68.41 Ветеринарна медицина: $T = \{a_1, \dots, a_{10}\}$. Апріорно виділені й 10 ключових термінів $R = \{b_1, \dots, b_{10}\}$, відповідно: b_1 – штамп, b_2 – фізіологічні порушення, b_3 – туберкульоз, b_4 – патологічний матеріал, b_5 – мікобактерії, b_6 – лікувальний препарат, b_7 – інфекційний процес, b_8 – ветеринарно-санітарна експертиза, b_9 – вароотоз бджіл, b_{10} – біпін. Графічну інтерпретацію предиката $P(t_i, r_j)$ наведеного прикладу при $1 \leq i \leq 10, 1 \leq j \leq 10$ показано на рис. 2.

Рис. 2. Графічна інтерпретація предиката $P(t, r)$

Розбиття на шари $\mathcal{G}_1$ множини текстів документів T

$$\mathcal{G}_1 = \{\{a_1, a_4, a_6\}, \{a_2, a_3, a_5\}, \{a_7, a_{10}\}, \{a_8, a_9\}\}.$$

У ролі множини T_1 , яка включає назви підтем, що об'єднують тексти документів множини T , виступає сукупність назв всіх шарів розбиття $\mathcal{G}_1$: $T_1 = \{\mu_1, \mu_2, \mu_3, \mu_4\}$, де $\mu_1 = \{a_1, a_4, a_6\}$, $\mu_2 = \{a_2, a_3, a_5\}$, $\mu_3 = \{a_7, a_{10}\}$, $\mu_4 = \{a_8, a_9\}$. Розбиття на шари $\mathcal{G}_2$ множини ключових понять R

$$\mathcal{G}_2 = \{\{b_1, b_2, b_4\}, \{b_3, b_7\}, \{b_5\}, \{b_6, b_8, b_9, b_{10}\}\}.$$

У ролі множини R_1 , що являє собою множину назв підрубрик ключових термінів множини R , виступає сукупність назв всіх шарів розбиття $\mathcal{G}_2$: $R_1 = \{\nu_1, \nu_2, \nu_3, \nu_4\}$, де $\nu_1 = \{b_1, b_2, b_4\}$, $\nu_2 = \{b_3, b_7\}$, $\nu_3 = \{b_5\}$, $\nu_4 = \{b_6, b_8, b_9, b_{10}\}$. Множина R_1 являє собою множину понять, що відображаються цими ключовими термінами, які можна визначити як дескриптори, що об'єднують ключові терміни за умовною і безумовною еквівалентністю. Зв'язок між назвами та шарами розбиття, що означаються ними, записується у вигляді предиката

$$F(r, \rho) = (r^{b_1} \vee r^{b_2} \vee r^{b_4})\rho^{\nu_1} \vee (r^{b_3} \vee r^{b_7})\rho^{\nu_2} \vee (r^{b_5})\rho^{\nu_3} \vee (r^{b_6} \vee r^{b_8} \vee r^{b_9} \vee r^{b_{10}})\rho^{\nu_4}.$$

Значеннями змінної ρ служать назви розбиття $\mathcal{G}_2$, які і є назвами

підрубрик, що включають ключові поняття r .

На етапі моделювання системи асоціації наукових і технічних понять та областей знань вводиться предикат аналітико-синтетичної обробки документа $Z(\tau, \rho) = \varepsilon$, $\varepsilon = \{0, 1\}$, який відображає відповідність ($\varepsilon = 1$) або невідповідність ($\varepsilon = 0$) предмета τ , що розглядається в тексті документа t , поняттю ρ , що виражається ключовим терміном r

$$P(t, r) = Z((g(t), f(r))) = Z(\tau, \rho),$$

де $\tau = g(t)$ – функція розуміння тексту, що відображає множину T на множину T_1 ; $\rho = f(r)$ – функція розуміння ключового слова, що відображає множину R на множину R_1 .

Запропоновано та розроблено математичні засоби опису дескрипторно-текстового предиката, функції розуміння тексту і функції розуміння лексичних одиниць, включаючи використання інструментальних засобів компараторної ідентифікації, необхідних для моделювання відношень між бібліотечними об'єктами.

Розроблено елементи архітектури галузевої інформаційної системи. Розглянуто операційний, технологічний, сервісний, системний рівні архітектури АІБС. Введено додатковий системний рівень, який описує функціональні сервіси АІБС. Визначено ключові концепції створення АІБС з огляду на сучасні тенденції розвитку інформаційних технологій. Запропоновано та розроблено моделі метаданих рубрикатора відповідно до стандарту ISO 19502 з використанням рекомендацій OMG MOF та мови моделювання UML. Моделі метаданих формується з наборів взаємопов'язаних моделей даних. В свою чергу моделі даних включають в себе ER-модель та реляційну модель сховища даних, модель збиру даних, модель доступу до даних та модель адміністрування. Моделі даних зв'язуються між собою завдяки правилам відображення, оскільки компоненти в одній моделі даних є похідними по відношенню до іншої моделі даних. Таким чином, повна модель метаданих складається із моделей як операційних, так і аналітичних даних з міжмодельним відображенням у середині та між операційними і аналітичними даними.

Четвертий розділ присвячено практичній реалізації одержаних результатів дисертаційного дослідження. Подано інформаційну технологію формування галузевих електронних ресурсів у ветеринарній медицині, описано впровадження результатів дослідження при створенні електронного рубрикатора з ветеринарної медицини, а також програмне забезпечення та супровід рубрикатора.

За основу розробки галузевого рубрикатора було взято розроблені моделі метаданих та структуру рівнів, яка представлена на рис. 3.

Рис. 3. Базова модель структури рівнів для розробки галузевого рубрикатора

Така побудова дозволила враховувати як вертикальні зв'язки в кожній окремій проблемі галузі, так і горизонтальні її зв'язки з іншими об'єктами. Вертикальні зв'язки розкривають поділ напрямків галузі в глибину, предметні та процесуальні фасети дають змогу розкривати не тільки об'єкти та методи, властивості, але й прослідкувати зв'язки з іншими проблемами та методами.

Структура рубрикатора поділяється на лінійну ієрархічну частину, яка відображає поділ науки, та фасетно-блочну частину, яка відображає взаємозв'язки між різними проблемами. Фасетно-блочна частина виражена в даному випадку як допоміжний апарат, в якому відображаються зв'язки та взаємозв'язки. Прикладом такої системи є рубрика 68.41.37 *Ветеринарна фармакологія*. Очевидно, що в рубриці повинен бути загальний матеріал щодо властивостей лікарських речовин та змін в організмі тварин під їх впливом, матеріал про окремі лікарські речовини:

Вплив лікарських речовин на збудників хвороб тварин див. 68.41.35 *Ветеринарна мікробіологія*

Токсичність лікарських речовин див. 68.41.39 *Ветеринарна токсикологія*

Гемотерапія тварин див. 68.41.43.31 *Ветеринарна гемотерапія*

Дієтотерапія тварин див. 68.41.43.33 *Ветеринарна дієтотерапія*

Тканинна терапія тварин див. 68.41.43.37 *Тканинна терапія в ветеринарії*

Техніка введення лікарських речовин тваринам та птиці див. 68.41.43.47 *Терапевтична техніка*.

Програма підтримки та ведення рубрикатора створена за допомогою вбудованої до пакету CDS/ISIS мови програмування PASCAL. Моделі, методи та розроблені у роботі математичні засоби, які було застосовано, дозволили в автоматизованому режимі створювати тематичні покажчики, індексувати та рубрикувати документи відповідно до наведеної схеми (рис. 4).

Рис. 4. Принципова схема роботи електронного рубрикатора

Найважливішим параметром використання рубрикатора є зміст інформаційних документів, який розподілено за принципами: об'єктний, спеціалізований, традиційний та предметний. Рубрикатор є прагматичною класифікацією, створюється на базі реальних інформаційних потоків і інформаційних потреб науковців і фахівців, та є предметом погодження бібліотек і органів інформації для виконання конкретних практичних задач у процесі обміну інформацією.

Оцінку ефективності використання рубрикатора проведено згідно з рекомендаціями міждержавного стандарту з інформації, бібліотечної та видавничої справи відповідно до таких критеріїв як: *precision* — коефіцієнт точності; *recall* — коефіцієнт повноти; *foraged* — коефіцієнт помилкової видачі;

fallout — коефіцієнт шуму; pertinence — коефіцієнт пертинентності, визначений як відношення кількості отриманої корисної інформації до інформаційної потреби і являє собою суб'єктивну міру.

Для визначення необхідної кількості документів у рубрикованій базі даних, до якої застосовано тестову пошукову вибірку, результати якої застосовані для розрахунку коефіцієнтів precision, recall, forged та fallout, використано засоби математичної статистики та теорії ймовірності. Загальна кількість оброблених документів T складає 300 примірників.

Виділено чотири класи інформаційних ресурсів по відношенню до включення у відповідь і релевантних рубриці документів: $n_{yy} = R_a$ — кількість документів, релевантних рубриці у пошуковій видачі; $n_{yn} = A - R_a$ — кількість документів що не є релевантними рубриці у пошуковій видачі; $n_{ny} = R - R_a$ — кількість не виданих у пошукову видачу, але релевантних рубриці документів; $n_{nn} = (T - R) - (A - R_a)$ — кількість не виданих у пошукову видачу та не релевантних рубриці документів, де $A = 63$ — кількість результатів видачі рубрикатора у відповідь на запит рубрики 68.41 «Ветеринарна медицина»; $R = 98$ — кількість документів, релевантних рубриці 68.41 «Ветеринарна медицина» у базі даних, відповідно до висновку експерта; $R_a = 59$ — кількість виданих релевантних рубриці документів відповідно до висновків експерта.

Тоді коефіцієнт точності видачі рубрикатора

$$precision = \frac{n_{yy}}{n_{yy} + n_{yn}} = \frac{R_a}{A} = \frac{59}{63} = 0,9365,$$

коефіцієнт повноти видачі рубрикатора

$$recall = \frac{n_{yy}}{n_{yy} + n_{ny}} = \frac{R_a}{R} = \frac{59}{98} = 0,602,$$

коефіцієнт помилкової видачі рубрикатора

$$forged = \frac{n_{yn}}{n_{nn} + n_{yn}} = \frac{A - R_a}{(T - R) - (A - R_a) + (A - R_a)} = \frac{63 - 59}{(300 - 98)} = 0,0198,$$

коефіцієнт шуму видачі рубрикатора

$$fallout = \frac{n_{yn}}{n_{yy} + n_{yn}} = \frac{A - R_a}{A} = \frac{63 - 59}{63} = 0,0635.$$

Для урахування балансу між повнотою та точністю було використано F -міру Ван Ризбергена, яка представляє собою середньо гармонічне коефіцієнтів повноти та точності

$$F = \frac{2 \times recall \times precision}{recall + precision} = \frac{2 \times 0,602 \times 0,9365}{0,602 + 0,9365} = \frac{1,1275}{1,5385} = 0,7329.$$

Таким чином, математичний апарат, використаний при розробці рубрикатора, задовольняє інформаційні потреби пошуку у контексті задачі «автоматична рубрикація» з суб'єктивним рівнем пертинентності, підтвердженим експертом предметної області, що складає не нижче 90%, та з мірою пошуку Ван Ризбергена, яка дорівнює 0,7329.

Розроблений рубрикатор впроваджений у Науковому інформаційно-бібліотечному центрі (НІБЦ) Інституту експериментальної та клінічної ветеринарної медицини (ІЕКВМ) (м. Харків) та в Харківській державній зооветеринарній академії (ХДЗВА) з програмним комплексом аналітичного розпису періодичних видань з ветеринарії та відображається як пошуковий словник в локальних мережах цих установ. Рубрикатор відображає структуру наукових досліджень; порівнює проблеми; розкриває всебічно кожен з них через її взаємозв'язки з іншими; дозволяє складати бібліографічні покажчики; надає можливість розподілити напрямки ветеринарної медицини між бібліотеками для створення розподіленої автоматизованої системи рубрикування. Розроблені рекомендації застосування рубрикатора запропоновані для використання в бібліотеках та інформаційних структурах ветеринарного профілю. Показано, що створена версія електронного рубрикатора може бути використана як складова частина електронного рубрикатора з питань сільського господарства.

Одержані результати практично реалізовані у ХДЗВА у вигляді: 1. Рекомендацій: "Рубрикатор з ветеринарної медицини. Структура, правила використання та ведення". 2. Моделей метаданих для формування фонду галузевих електронних бібліотечних ресурсів. 3. Автономної системи автоматичного рубрикування та аналітичного розпису. 4. Компонента програмного продукту AIBC LiberMedia. 5. Компонента програмного продукту бібліотечної пошукової машини на Web-сервері ХДЗВА. 6. Методичного забезпечення та БД нормативно-правових документів і комп'ютерної програми навчального процесу для студентів факультету ветеринарної медицини, технологічного факультету та факультету менеджменту. 7. Комп'ютерної версії з навчального процесу з екології та ділової гри з екологічного моніторингу.

ВИСНОВКИ

У дисертаційній роботі вирішена актуальна науково-практична задача підвищення ефективності формування галузевих електронних бібліотечних ресурсів на основі моделей, методів та інформаційних технологій інтелектуальної обробки документів у бібліотеках.

В процесі виконання роботи отримані наступні результати:

1. Досліджено логіко-семантичну структуру предметної області формування фондів галузевих бібліотек, визначено особливості їхньої

автоматизації та застосування математичних моделей процесу аналітико-синтетичної обробки текстів документів. Розглянуто методи формування галузевих електронних ресурсів. Обґрунтовано використання методу компараторної ідентифікації, алгебри скінченних предикатів та теорії категорій для моделювання процесів обробки текстових даних у документах спеціалізованих фондів.

2. Побудовано модель ідентифікації змістовних ознак інформаційних лінгвістичних об'єктів, що входять до складу галузевих фондів, розроблено математичну модель інтелектуальної обробки даних, яка дозволяє перейти до знання-орієнтованих методів формування фондів. Вирішена задача створення та поповнення галузевих термінологічних словників на основі застосування методу компараторної ідентифікації, теорії категорій та логічних мереж.

3. Розроблено засоби автоматичної побудови семантичної мережі спеціалізованого україномовного тезаурусу з ветеринарної медицини, систематизації текстів повнотекстової бази даних і пошуку в множині текстів, релевантних запиту користувача, вираженому обмеженою природною мовою. Розроблено формальні засоби компараторної ідентифікації для моделювання аналітико-синтетичної обробки бібліотечних об'єктів: дескрипторно-текстовий предикат, предикат аналітико-синтетичної обробки документа, описано їх властивості, потрібні для опису функції розуміння тексту і лексичної одиниці, що виражає певне поняття.

4. Розроблено прикладну інформаційну технологію інтеграції розподілених галузевих ресурсів, реалізовано розроблені методи і моделі у вигляді модулю, інтегрованого в підсистему каталогізації і користувацького доступу існуючої АІБС; розроблено проблемно-орієнтовані засоби навігації в галузевих інформаційних масивах. Визначено ключові концепції та рівні архітектури інформаційної бібліотечної системи. Розроблено модель метаданих для формування фонду галузевих електронних бібліотечних ресурсів, що дозволяє автоматизувати класифікацію і сегментацію цифрових ресурсів.

5. Розроблено електронний галузевий рубрикатор з ветеринарної медицини, його структуру, правила використання та супроводження, який відповідає сучасному розвитку ветеринарної науки та лексичним особливостям української мови, що дозволило удосконалити довідково-пошукову функцію, розподіл напрямків предметної області між ветеринарними установами, та об'єднання електронних документно-інформаційних ресурсів.

6. Практичні результати дисертаційних досліджень впроваджено у науково-методичному відділі наукової бібліотеки Харківського національного медичного університету, що підвищило рівень автоматизації аналітико-синтетичної обробки підфону спеціалізованих текстів, та у науковому інформаційно-бібліотечному центрі Харківської державної зооветеринарної академії, що дозволило поліпшити якість формування галузевих ветеринарних фондів та підвищити релевантність і повноту обслуговування читачів.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Гончаров С.В. Ключові концепції архітектури цифрової бібліотеки та її базові компоненти / С.В.Гончаров – Вісник Книжкової палати – К.: Книжкова палата України, 2006 – № 1 (114) – С. 35-37.

2. Гончаров С.В. Створення та поповнення електронних галузевих термінологічних словників / С.В. Гончаров, Н.В. Борисова – Вестник Херсонского национального технического университета. – Херсон: ХНТУ, 2009. – № 1 (34). – С. 204-206.

Здобувачем запропоновано використання алгебро-логічних методів для формалізації пошуку нових слів та поповнення електронних галузевих термінологічних словників.

3. Гончаров С. В. Использование предикатных категорий для представления информации в документах / С. В. Гончаров, Сайед Мохаммад Таухид Сиддики – Вісник Національного технічного університету "Харківський політехнічний інститут". – Харків: НТУ "ХПІ". – 2009 – № 4 – С. 96-102.

Здобувач на основі аналізу існуючих моделей запропонував спосіб ефективного аналізу текстової інформації за допомогою методу компараторної ідентифікації та дескрипторно-текстового предикату.

4. Гончаров С.В. Інформаційна технологія формування галузевих електронних ресурсів у ветеринарній медицині / Н.В.Шаронова, С.В. Гончаров, Н.В. Борисова. – Проблеми інформаційних технологій. – Херсон: ХНТУ, 2008. – № 02 (004). – С. 115-120.

Здобувачем запропоновані математичні моделі формування галузевих електронних ресурсів у прикладній галузі ветеринарної медицини.

5. Гончаров С.В. Алгебра скінченних предикатів як складова інформаційних технологій / С. В. Гончаров, О. В. Каніщева. – Збірник наукових праць Військового інституту Київського нац. ун-ту ім. Тараса Шевченка, – К.: ВІКНУ, 2009, – Вип. №22. – С. 80-85.

Здобувач запропонував підхід до використання алгебри скінченних предикатів, методу компараторної ідентифікації як одного з логічних методів інтелектуальної обробки даних.

6. Гончаров С.В. До проблеми створення корпоративної регіональної мережі / Г.В. Шемаєва, С.В. Гончаров, Т.Н. Приходько. – Інформаційні ресурси та їх використання в агропромисловому виробництві: зб. наук. праць Інституту аграрної економіки, № 2. – К., 2000. – С.177-180.

Здобувачем запропоновано підхід до визначення архітектури корпоративної регіональної мережі.

7. Гончаров С.В. Використання предикатних категорій для формалізації знань, представлених за допомогою логічних моделей – Проблеми інформаційних технологій. – Херсон : ХНТУ, 2009. – № 01 (005). – С. 47-50.

Здобувачем запропоновано використання предикатних категорій для формалізації знань, представлених за допомогою логічних моделей.

8. Гончаров С. В. Применение предикатных категорий для формализации знаний / С. В. Гончаров, Сайед Мохаммад Таухид Сиддики // 13-ий

Міжнародний молодіжний форум "Радіоелектроніка і молодь в ХХІ ст.". – Харків: ХНУРЕ, 2009. – С. 138.

Здобувачем запропоновано математичну модель ідентифікації документів у системах формалізації знань.

АНОТАЦІЇ

Гончаров С.В. - Методи та інформаційні технології формування галузевих електронних бібліотечних ресурсів. – Рукопис.

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

Дисертація присвячена дослідженню та розробці методів, моделей та інформаційних технологій формування галузевих електронних бібліотечних ресурсів, які ґрунтуються на методах інтелектуальної обробки даних. У роботі запропоновано оригінальний метод формування галузевих фондів інформаційно-бібліотечних систем, заснований на застосуванні методу компараторної ідентифікації, теорії категорій та апарату логічних мереж, що дозволило значно підвищити ступінь автоматизації формування галузевих бібліотечних фондів за рахунок визначення та ефективної обробки текстів галузевої спрямованості. Набули подальшого розвитку моделі та інформаційна технологія обробки інформаційних об'єктів ветеринарної галузевої бібліотеки: розроблено алгоритми і реалізовано процедури автоматичної індексації і предметизації сукупності галузевих текстів в автоматизованих інформаційних бібліотечних системах (АІБС), автоматизовано процес створення та поповнення галузевих термінологічних словників. Удосконалено процеси аналітико-синтетичної обробки підмножини спеціалізованих текстів в АІБС, що дозволило реалізувати методи систематизації та рубрикації текстових документів певної тематики, автоматичної побудови семантичної мережі багатомовного тезаурусу для реферування, індексації, динамічної класифікації предметної галузі комплексу ветеринарних дисциплін. Розроблено рекомендації щодо застосування методу формалізації структури міждисциплінарної галузі комплексу ветеринарних дисциплін до інших міждисциплінарних галузей наукового знання.

Ключові слова: інформаційні технології, галузеві електронні ресурси, теорія категорій, компараторна ідентифікація, інтелектуальна обробка даних, логічні мережі, рубрикатор.

Гончаров С.В. Методы и информационные технологии формирования отраслевых электронных библиотечных ресурсов. – Рукопись.

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

технический университет "Харьковский политехнический институт", Харьков, 2010.

Диссертация посвящена исследованию и разработке методов, моделей и информационных технологий формирования отраслевых электронных библиотечных ресурсов, основанных на методах интеллектуальной обработки данных.

Решению указанной задачи в работе предшествует анализ современных подходов к автоматизации библиотечной деятельности, существующих проблем создания отраслевых фондов, связанных с интеллектуальной обработкой документов. Анализ существующих средств интеллектуальной обработки информации показал отсутствие единого математического аппарата для эффективного моделирования процессов библиотечной деятельности, необходимость построения математических моделей формирования отраслевых фондов. На основании этого анализа в работе предложен оригинальный метод формирования фондов, основанный на использовании метода компараторной идентификации, теории категорий и аппарата логических сетей, что позволило значительно повысить уровень автоматизации формирования отраслевых библиотечных фондов за счет учета особенностей автоматизированных информационных библиотечных систем (АИБС), идентификации и эффективной обработки текстов отраслевой направленности.

В работе получили дальнейшее развитие модели и информационная технология обработки информационных объектов ветеринарной отраслевой библиотеки, а именно: разработаны алгоритмы и реализованы процедуры автоматической индексации и предметизации совокупности отраслевых текстов в автоматизированных информационных библиотечных системах с учетом стилистических особенностей их графемной структуры, идентификации их позиции в иерархической структуре динамической классификационной системы, объективно отображающей подмножество тем предметной области, что позволило автоматизировать процесс создания и пополнения отраслевых терминологических словарей.

Усовершенствованы процессы аналитико-синтетической обработки подмножества специализированных текстов в АИБС, что позволило реализовать методы автоматического построения семантической сети украиноязычного тезауруса для реферирования, индексации, построения динамической классификации предметной области комплекса ветеринарных дисциплин, систематизации и рубрикации текстовых документов. Разработаны рекомендации по использованию метода формализации структуры междисциплинарной области комплекса ветеринарных дисциплин по отношению к другим междисциплинарным областям научного знания.

Разработана электронная версия рубрикатора, соответствующая современному развитию ветеринарной науки и лексическим особенностям украинского языка, что позволило усовершенствовать справочно-поисковую функцию, распределение направлений предметной области между ветеринарными организациями, и объединение электронных документно-информационных ресурсов.

Практические результаты исследований внедрены в научно-методическом отделе научной библиотеки Харьковского национального университета и в научном информационно-библиотечном центре Харьковской государственной зооветеринарной академии, что позволило улучшить качество формирования отраслевых ветеринарных фондов и повысить релевантность и полноту обслуживания пользователей.

Ключевые слова: информационные технологии, отраслевые электронные ресурсы, теория категорий, компараторная идентификация, интеллектуальная обработка данных, логические сети, рубрикатор.

Goncharov S.V. Methods and information technologies of specialized electronic libraries resources formation. – *Manuscript.*

Thesis for a candidate degree in technical science, speciality 05.13.06 – information technologies. – National Technical University “Kharkiv Polytechnic Institute.” – Kharkiv, 2010.

This dissertation establishes and analyzes methods, models and information technologies of specialized electronic library resources formation, based on the methods of intelligent database. The thesis suggests a special method of formation of information library system specialized funds based on the method of comparator identification, on the theory of categories and on the logical nets apparatus, which allowed to significantly increase the level of automation of specialized library funds formation due to identification and efficient processing of the texts of a certain field of knowledge. The models and information technologies of processing of information objects of veterinary libraries have been further developed: algorithms have been made up, the procedures of automatic indexation and subject analysis of specialized texts in Automatic Information Library Systems have been realized, the processes of specialized dictionaries compiling and expanding have been automated. Analytic-synthetical processing of the specialized texts subset in Automatic Information Library Systems has been updated, which allowed to apply the methods of systematization and rubrication of specialized text documents, to automatically build a semantic network of a multilingual thesaurus for abstracting, indexation and dynamic classification within the subject field of veterinary. The paper provides recommendations as to applying the method of structure formalization of interdisciplinary sphere of the complex of veterinary disciplines to other interdisciplinary spheres of scientific knowledge.

Key words: information technologies, specialized electronic resources, the theory of categories, comparator identification, intelligent database, logical nets, rubricator.