ХАРКІВСЬКИЙ ДЕРЖАВНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
РАДІОЕЛЕКТРОНІКИ

ХАЙРОВА Ніна Феліксівна

УДК 681.518:519.767

РОЗРОБКА МАТЕМАТИЧНОГО І ЛІНГВІСТИЧНОГО ЗАБЕЗПЕЧЕННЯ АВТОМАТИЗОВАНИХ ІНФОРМАЦІЙНО-БІБЛІОТЕЧНИХ СИСТЕМ

05.13.06 – автоматизовані системи управління
та прогресивні інформаційні технології

Автореферат дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2000

Дисертацією є рукопис.

Робота виконана в Харківському гуманітарному інституті “Народна українська академія” Міністерства освіти і науки України.

Науковий керівник – доктор технічних наук, професор Шаронова Наталія Валеріївна, Харківський гуманітарний інститут “Народна українська академія”, проректор з наукової роботи

Офіційні опоненти: – доктор технічних наук, професор Петров Едуард Георгійович, Харківський державний технічний університет радіоелектроніки, завідувач кафедрою системотехніки;

– доктор технічних наук, професор Сенченко Микола Іванович, Книжкова палата України, м. Київ, директор

Провідна установа – Харківський державний політехнічний університет, кафедра автоматизованих систем управління, Міністерства освіти і науки України, м. Харків.

Захист відбудеться “ 26 ” квітня 2000 р. о 13 годині на засіданні спеціалізованої вченої ради Д 64.052.01 в Харківському державному технічному університеті радіоелектроніки, за адресою 61166, м.Харків, просп.Леніна, 14; т. 433-053.

З дисертацією можна ознайомитись у бібліотеці Харківського державного технічного університету радіоелектроніки, просп.Леніна, 14.

Автореферат розісланий “ 21 ” березня 2000 р.

Вчений секретар
спеціалізованої вченої ради Саенко В.И.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. На порозі XXI століття інформація стає одним із найбільш значущих стратегічних ресурсів, які мають вирішальний вплив на розвиток суспільства. У Законі України "Про національну програму інформатизації" підкреслюється роль бібліотеки і визначаються її завдання в процесі формування нового інформаційного суспільства. З розвитком глобальної інформаційної мережі Internet, з появою нових нетрадиційних джерел інформації, в тому числі повнотекстових електронних баз даних, бібліотека переростає з зберігача в розробника інформаційних ресурсів і провідника в світовому інформаційному середовищі. При цьому автоматизація всіх процесів бібліотечної діяльності повинна забезпечити доступ користувача до електронного каталога та інших баз данних бібліотеки з максимально адекватною відповідністю отриманої інформації до читацьких запитів. Сьогоднішня автоматизована інформаційно-бібліотечна система (АІБС) являє собою найбільш передову сучасну інформаційно-пошукову систему, що забезпечує пошук серед великих масивів інформації за різними критеріями.

Великий внесок в розробку теоретичних і прикладних питань автоматизації інтегрованої бібліотечної системи і окремих її функціональних модулів внесли такі видатні вчені як Ф.Воройський, І.Коровякова, Робін Т. Гарбор, М.Сенченко, Дж.Солтон, М.Селтон, Л.Філіппова, І.Фоменко, Дж.Хеклі, В.Цуркан, Я.Шрайберг та ін.

Однією з головних вимог, що пред'являються сьогодні до АІБС, є забезпечення користувача в реальний проміжок часу повною і релевантною інформацією, що можливо тільки при наближенні інформаційно-пошукових запитів до природної мови. Важливою складовою частиною досліджень, які проводяться в даному напрямку є розробка математичного і лінгвістичного забезпечення АІБС, що являють собою моделі, алгоритми і методи, які охоплюють процеси класифікації, предметизації, реферування, анотування і т.п. При цьому ставиться завдання розв'язання трьох теоретичних проблем: подання знань бібліотечної системи; комп'ютерна лінгвістика, вирішення якої забезпечить розробку природномовного інтелектуального інтерфейсу бібліотечної системи і моделювання інтелекту людини в процесі "розуміння" під час аналітико-синтетичної обробки документа.

Найбільш перспективним сьогодні стає використання моделей і методів інформаційних технологій, що базується на результатах, отриманих при розв'язанні проблем штучного інтелекту (ШІ). Наука, що вивчає механізми природного інтелекту з метою використання набутих знань для створення систем штучного інтелекту, розробляється науковою школою проф. Ю.П. Шабанова-Кушнаренко і носить назву теорії

інтелекту. Мета робіт, присвячених ШІ, є створення комп'ютерних систем, що автоматизують інтелектуальну діяльність людини. Істотний внесок в розв'язання питань моделювання розуміння в інтелектуальних системах внесли Т.Віноград, Н. Леонова, Д.Мінський, С. Осуга, О. Перевозчикова, Е.Попов, Д.Поспелов, Ю.Саєкі, К. Філлмор, Р.Шенк, К. Ющенко та ін.

Під час розробки математичного забезпечення АІБС необхідний повний, однозначний і експліцитний опис процесів аналітико-синтетичної обробки, який призначений для ЕОМ і може зробити доступними для них властиві людині операції з обробки текстової інформації. Незважаючи на досягнуті результати в області моделювання інтелекту, питання семантичної обробки текстової інформації все ще недостатньо вивчені. Досвід дослідження і моделювання бібліотечних процесів довів необхідність серйозної розробки лінгвістичного забезпечення, який включає лінгвістичний процесор (що дозволяє передусім трансформувати пошукові запити користувача, виражені природною мовою, в інформаційну мову системи), різного роду класифікаційні схеми, засоби предметного пошуку, рубрикатори, тезауруси і т.д.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконувалася на кафедрі інформаційних технологій і документознавства Харківського гуманітарного інституту "Народна українська академія" відповідно до плану науково-дослідної роботи, в рамках розробки комплексної наукової теми кафедри "Дослідження актуальних проблем побудови інтелектуальних систем". Розроблені в дисертації алгоритми були використані під час виконання держбюджетної теми Міністерства оборони "Корекція-А" Харківського військового університету (відповідно до договору від 19.10.98).

Мета і задачі дослідження. Метою дисертаційної роботи є дослідження і розвиток лінгвістичного і математичного забезпечення автоматизованих інформаційно-бібліотечних систем; розробка нових інформаційних технологій, що базуються на формалізації і моделюванні інтелекту в процесі семантичної обробки текстів документів і тісно пов'язаною з нею предметно-тематичною асоціацією користувача в процесі організації тематичного пошуку; їх реалізація у вигляді алгоритмів і програм підсистем АІБС.

Для досягнення поставленої мети необхідно вирішити такі завдання:

1. Розробити метод класифікації документальної інформації, що базується на моделюванні системи асоціацій наукових і технічних понять та дозволяє підвищити релевантность та повноту інформації, що видається АІБС.

2. Розробити інструментальні засоби компараторної ідентифікації, необхідні для моделювання відносин бібліотечних об'єктів.

3. Розробити математичні засоби опису дескрипторно-текстового предиката, опису функцій розуміння тексту і лексичних одиниць, що виражає теми (поняття, рубрики).

4. Побудувати математичні моделі відносин об'єктів на графемному і семантико-синтаксичному рівнях лінгвістичного процесора автоматизованої інформаційно-бібліотечної системи.

5. Сформулювати і дослідити найбільш важливі задачі автоматизованого аналізу текстової інформації, які використьвуються під час розробки систем аналітико-синтетичної обробки документальної інформації в АІБС, розробити методи і алгоритми рішення цих завдань і реалізувати їх у вигляді програмних систем обробки документів повнотекстових баз даних.

Наукова новизна одержаних результатів. У процесі розв'язання завдань, відповідно до мети роботи отримано такі результати:

- в роботі вперше запропоновано і обгрунтовано використання методу компараторної ідентифікації для моделювання процедур аналітико-синтетичної обробки текстів документів, які циркулюють в АІБС. Введено дескрипторно-текстовий предикат, функції розуміння тексту і ключових термінів, що дозволило розробити математичне забезпечення процедур каталогізації, систематизації і предметизації, безпосередньо пов'язаних із організацією тематичного бібліотечного пошуку;
- вдосконалено методику формалізації інтелектуальної діяльності людини по розумінню і класифікації за змістовими ознаками лексичних одиниць (ЛО), що дозволило розробити концептуальну модель динамічного рубрикатора;
- набула подальшого розвитку розробка дескрипторної мови з лінійною (позиційно-дужковою) граматикою, яка відображає взаємне розміщення лексичних одиниць і фрагментів документа;
- вперше запропоновано і розроблено метод контекстного аналізу текстів повнотекстової БД АІБС, який дозволяє зняти частину мовної багатозначності;
- розроблено алгоритм, реалізований у вигляді програмного комплексу Texs, що подає динамічну підсистему систематизації і предметизації АІБС, яка дозволяє переглядати структуру класів документів фонду в сучасний момент.

Практичне значення одержаних результатів. Розроблене в дисертаційній роботі математичне і лінгвістичне забезпечення орієнтоване на автоматизацію аналізу змісту документів повнотекстової бази даних, тобто приписуванні накопиченим одиницям інформації позначень, які адекватно відображають їх зміст, автоматизації аналітико-синтетичної обробки документа (класифікації, предметизації, реферуванні, що є найбільш трудомісткими бібліотечними процесами).

Результати дисертаційних досліджень були використані як частина програми із автоматизованої аналітико-синтетичної обробки текстів українською та російською мовами в науково-методичному відділі Центральної наукової бібліотеки Харківського національного університету ім. В.Н. Каразіна.

Наукові положення, висновки і рекомендації дисертаційної роботи використані в навчальному процесі при підготовці курсів "Машинний переклад" і "Автоматичне реферування" для студентів старших курсів спеціальності 7.030.504 "Прикладна лінгвістика" на кафедрі інформаційних технологій і документознавства Харківського гуманітарного інституту "Народна українська академія". Результати дисертаційних досліджень були використані у посібнику з курсу “Машинний переклад”, що отримав гриф навчального посібника Міністерства освіти України і диплом виставки-ярмарки науково-педагогічних ідей "Освіта Харківщини" за 1998 рік.

Теоретичні і практичні результати дисертаційної роботи були використані також під час виконання науково-дослідних робіт в Харківському військовому університеті по темі "Корекція-А" (договір від 19.10.98).

Особистий внесок здобувача. Всі результати дисертації отримані автором самостійно. У роботі [1], яка виконана в співавторстві, дисертанту належить розробка морфологічного і синтаксичного етапів аналізу лінгвістичного процесора систем машинного перекладу. У працях [3, 6] автором розроблені постановка задачі та інструментальні засоби компараторної ідентифікації, необхідні для моделювання процедур класифікації, предметизації і систематизації документів. У роботі [4] автору належить постановка задачі і розробка математичної моделі, що дозволяє зняти значну частину морфологічної омонімії на етапі контекстного аналізу. У працях [5, 12] здобувачем розроблені практичні рекомендації із використання методу компараторної ідентифікації для звуження семантичного поля текстів первинних документів в системах автоматичного реферування. У роботі [13] автору належить розробка алгоритму автоматичного рубрикатора і його програмна реалізація.

Апробація результатів дисертації. Основні положення і результати дисертаційної роботи були подані і розглянуті на:

– III міжнародному семінарі "Актуальні питання впровадження інформаційних технологій у документально-комунікаційній сфері", Харків, 1996 р.;

– Міжнародній конференції "Создание, интеграция, использование информационных ресурсов инновационного развития", Київ, 1997 р.;

– III міжнародній науково-методичній конференції "Досвід і проблеми реалізації ступеневої системи підготовки фахівців", Суми, 1997 р.;

– IV міжнародній науковій студентській конференції "Актуальные проблемы гуманитарных наук и их информационное обеспечение", Харків, 1997р.;

– II молодіжному форумі "Радиоэлектроника и молодежь в XXI веке", Харків, 1998 р.;

– Науково-методичній конференції "Підвищення ефективності навчального процесу у технічному закладі на базі засобів Multimedia", Харків, 1998 р.;

– II науково-методичній конференції "Використання комп'ютерних технологій у навчальному процесі", Харків, 1998 р.;

– V міжнародній конференції "Теория и техника передачи, приема и обработки информации" ("Телекоммуникации, радиотехника, электроника"), Судак, 1999 р.;

– республіканському науково-практичному семінарі НАН України "Системный анализ, математическое моделирование и принятие решений в социально-экономических и технических системах", Харків, 1999 р.

Публікації. Основні положення дисертації викладені в 13 друкованих роботах, з них – 1 навчальний посібник з грифом Міністерства освіти України, 6 статей в наукових журналах, 5 тез доповідей і одна депонова стаття.

Структура і об'єм роботи. Дисертаційна робота складається із вступу, чотирьох розділів, висновку, списку літературних джерел із 122 найменувань, чотирьох додатків; включає 18 малюнків, 4 таблиці. Загальний об'єм роботи становить 159 сторінок, в тому числі 124 сторінки основного тексту.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обгрунтована актуальність дисертаційної роботи, сформульовані основна мета і завдання досліджень, наведені відомості про зв'язки обраного напрямку досліджень із планами організації, де виконана робота. Дано стислу анотацію отриманих у дисертації рішень, відзначена їх практична цінність, приведені дані про використання результатів проведенних досліджень у народному господарстві.

У першому розділі проведено аналіз сучасного стану автоматизованих інформаційно-бібліотечних систем (АІБС), перспектив і проблем автоматизації бібліотечних процесів. Показано необхідність автоматизації процесів семантичної обробки документів: класифікації, предметизації, систематизації, тобто тих процесів, на яких засновується організація інформаційного обслуговання користувачів бібліотечних систем.

У розділі висвітлюється недостатня ефективність існуючих засобів семантичного (тематичного і предметного) пошуку в АІБС, що забезпечують на сьогодні низькі коефіцієнти повноти і точності. Тоді як релевантність інформації, знайденої АІБС, до запиту користувача значною мірою залежить від правил структурного об'єднання баз даних, в основі яких лежить інструмент (методика) класифікації об'єктів. Ключовою проблемою в цій області залишається автоматизація аналізу змісту текстів документів, тобто приписування накопиченим одиницям інформації позначень, які адекватно відображають їх зміст, так звана аналітико-синтетична обробка. Спираючись на специфіку документальних бібліотечних систем, в розділі доводиться провідна роль рубрикаторів і тезаурусів під час моделювання об'єктів і зв'язків, істотних для завдань аналітико-синтетичної обробки і пошуку документів.

У розділі обгрунтовується необхідність обліку аспектів природної мови (ПМ) під час організації вузько-тематичного пошуку в інформаційних системах, які працюють із повнотекстовими базами даних (full text system). Як семантично найбільш сильні і здатні забезпечити високу якість пошуку для основної маси запитів, пропонується використати дескрипторні мови з лінійною (позиційно-дужковою) граматикою, яка відображає взаємне розташування лексичних одиниць і фрагментів документа.

Для налагодження зв'язків між споживачами і відповідними документами повинні бути використані різноманітні лінгвістичні засоби обробки інформації, що дозволяють наблизити інтерфейс автоматизованих бібліотечних систем до ПМ. Створення такого інтерфейсу АІБС дозволить, передусім, трансформувати пошукові запити, виражені на ПМ, в інформаційну мову системи. У розділі показано актуальність даної проблеми, необхідність розробки нових алгоритмів і сформовано завдання, які необхідно вирішити в зв'язку з цим. При цьому основним призначенням лінгвістичного забезпечення в документальній підсистемі АІБС є розробка моделей графемного, морфологічного і семантико-синтаксичного етапів аналізу лінгвістичного процесора АІБС.

У даному розділі обгрунтовуються переваги методу компараторної ідентифікації під час моделювання процесів інтелектуальної обробки бібліотечних об'єктів.

Отримані такі висновки:

1. Сучасні бібліотечні системи повинні настроюватися на використання БД, що наближаються до експертних систем, до систем автоматизації інтелектуальної діяльності, які володіють здібністю навчання.

2. Під час розробки методики вузько-тематичного багатоаспектного пошуку релевантної інформації в БД АІБС необхідно використати модель предметної області (ПО), що являє собою рубрикатор, який автоматично настроюється на нову семантичну область.

3. Для розробки процедур динамічної класифікації і систематизації документів необхідно моделювати систему асоціацій наукових і технічних понять і областей, наблизивши тим самим дану проблему до області формалізації інтелекту.

4. Під час організації вузько-тематичного пошуку в інформаційних системах, які працюють з повнотекстовими базами даних (full text system) необхідно враховувати не лише логічний зв'язок термінів і понять, але і аспекти природної мови.

У відповівдності з цим метою роботи є дослідження і розвиток лінгвістичного і математичного забезпечення автоматизованих інформаційно-бібліотечних систем, що базуються на формалізації і моделюванні інтелекту в процесі семантичної обробки текстів документів; їх реалізація у вигляді алгоритмів і програм підсистем АІБС.

У другому розділі розроблені формальні засоби компараторної ідентифікації, необхідні для моделювання аналітико-синтетичної обробки бібліотечних об'єктів: математичні засоби опису дескрипторно-текстового предиката, предиката інтелектуальної аналітико-синтетичної обробки документа та їх властивостей, опису функції розуміння тексту і функції розуміння лексичної одиниці, що виражає певне поняття.

Базовими при використанні методу компараторної ідентифікації документів в повнотекстовій базі даних є дві множини: множина документів $T=\{t_i\}$, $1 \le i \le n$, що являє собою деяку, досить чітко окреслену, сукупність текстів повнотекстової бази даних і досить чітко окреслена множина ключових термінів $R=\{r_j\}$, $1 \le j \le m$. Розглядаючи всі можливі пари з множини $T \times R$,, компаратор формує предикат P, який задає відносини між текстами документів і ключовими термінами, що відображають зміст цих документів. Дескрипторно-текстовий предикат $P(t_k,r_q)$, що відображає відносини між елементами кожної пари t_k,r_q, представлений так:

$$P(t_k, r_q) = \varepsilon, \text{ де } t_k \in T, r_q \in R, \varepsilon = \{0,1\}. \quad (1)$$

Два тексти t і t' відносяться до однієї підтеми $(t, t' \in T)$, $t \sim t'$ тоді і тільки тоді, коли для $\forall r$: $P(t, r) = P(t', r)$. Два ключових поняття r і r′ відносяться до однієї підрубрики $(r,r' \in R)$, $r \sim r'$ тоді і тільки тоді, коли для $\forall t$: $P(t, r) = P(t, r')$.

Предикати еквівалентності E_1, заданий на множині $T \times T$, і E_2, заданий на декартовому добутку $R \times R$, що однозначно визначаються предикатом P, відображають відповідність текстів документів одній підтемі:

$$E_1(t_1, t_2) = \forall r \in R (P(t_1,r) \sim P(t_2,r)) \quad (2)$$

і відповідність ключових понять одній підрубриці:

$$E_2(r_1, r_2) = \forall t \in T\,(P(t,r_1) \sim P(t,r_2)). \quad (3)$$

Предикат $E_1(t_1, t_2)$ використовується для об'єктивного визначення відношення будь-яких двох текстів документів t_1 і t_2,, що належать множині T, до однієї підтеми. Предикат $E_2(r_1, r_2)$ можна використати для визначення відповідності будь-яких двох ключових понять, що належать множині R, єдиній підрубриці. Предикат E_1 визначає розбиття ϑ_1 множини T на шари текстів документів. Всі документи, що належать одному шару розбиття, відносяться до однієї підтеми. Будь-які документи, взяті з різних шарів розбиття, відносяться до різних підтем. Предикат E_2 визначає розбиття ϑ_2 множин*i* R на шари ключових понять, розподіляючи ключові поняття на підрубрики. Всі ключові поняття, що належать одному шару розбиття, відносяться до єдиної підрубрики, будь-які два ключових поняття, взяті з різних шарів розбиття ϑ_2, відносяться до різних підрубрик.

Розподіл текстів на підтеми і ключових понять на підрубрики можна виразити через предикат P, що об'єктивно визначається компаратором. Класу L_a всіх текстів $t \in T$, що відносяться до однієї підтеми, що містить текст $a \in T$, відповідає предикат $L_a(t)$:

$$L_a(t) = E_1(t, a) = \forall r \in R\,(P(t,r) \sim P(a,r)). \quad (4)$$

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

$$Q_b(r) = E_2(r, b) = \forall t \in T\,(P(t, r) \sim P(t, b)). \quad (5)$$

У роботі розглянуто приклад поділу на підкласи множини ключових термінів, що відносяться до предметної області комп'ютерних технологій і Internet. Графічна інтерпретація предиката $P(t_i, r_j)$, наведеного прикладу, при $1 \le i \le 10$, $1 \le j \le 10$, показана на рисунку 1.

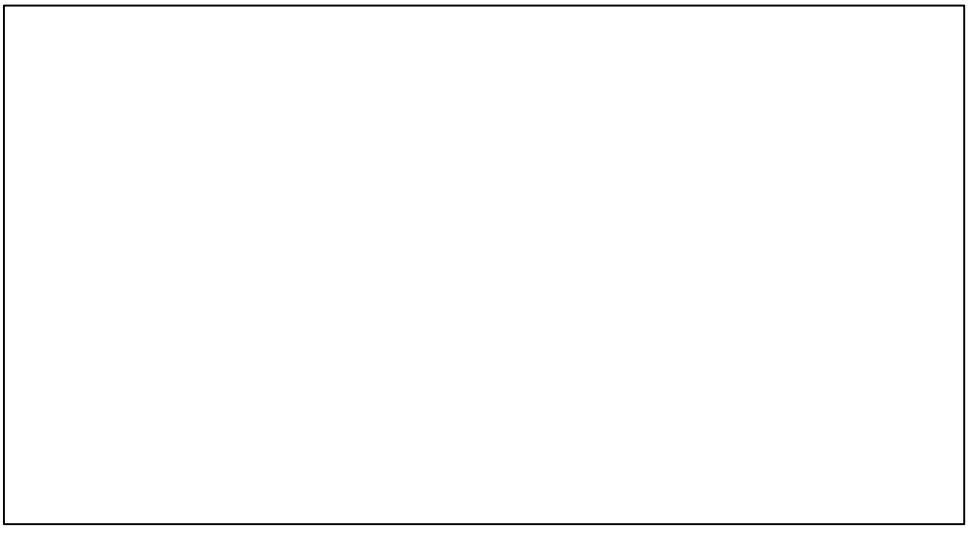


Рис. 1. Графічна інтерпретація предиката P(t, r).

Розбиття на шари ϑ_1 множини текстів документів T:

$$\vartheta_1=\{ \{ a_1, a_4, a_6\}, \{ a_2, a_3, a_5\}, \{ a_7, a_{10}\}, \{ a_8, a_9\} \}.$$

(6)

У ролі множини T_1, яка включає назви підтем, що об'єднують тексти документів множини T, виступає сукупність назв всіх шарів розбиття ϑ_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 ϑ_2 множини ключових понять R:

$$\vartheta_2=\{ \{ b_1, b_2, b_4\}, \{ b_3, b_7\}, \{ b_5 \}, \{ b_6, b_8, b_9, b_{10}\} \}.$$

(7)

У ролі множини R_1, що являє собою множину назв підрубрик ключових термінів множини R, виступає сукупність назв всіх шарів розбиття θ_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 їми, записується у вигляді предиката:

$$F(r,\rho) = (r^{b1} V r^{b2} V r^{b4}) \rho^{\nu1} V(r^{b3} V r^{b7}) \rho^{\nu2} V(r^{b5}) \rho^{\nu3} V(r^{b6} V r^{b8} V r^{b9} V r^{b10}) \rho^{\nu4}. \quad (8)$$

Значеннями змінної ρ служать назви розбиття ϑ_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 не відповідність $(\varepsilon=0)$ предмета τ, що розглядається в тексті документа t, поняттю ρ, що виражається ключовим терміном r.

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

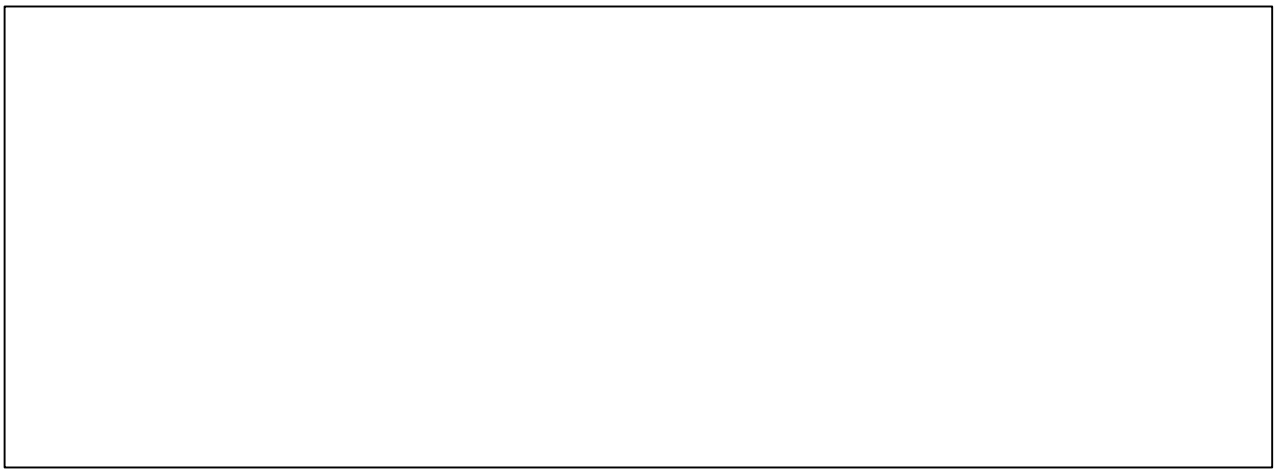
(9)

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.

Таким чином, у другому розділі дисертаційного дослідження розроблено математичні засоби опису дескрипторно-текстового предиката, функції розуміння тексту і функції розуміння лексичних одиниць, подаючи інструментальні засоби компараторної ідентифікації необхідні для моделювання відносин бібліотечних об'єктів.

У третьому розділі розглядаються моделі і алгоритми лінгвістичного забезпечення АІБС, що дозволяють істотно полегшити доступ користувача до інформаційних ресурсів бібліотеки.

Лінгвістична обробка текстів в АІБС, що включає формалізацію всіх рівнів мовної системи, подана схемою лінгвістичного процесора (ЛП):



На етапі графемного аналізу, подаючи текст природною мовою як цілісний об'єкт, елементами якого є знаки, організовані певним чином в рядки: ТЕКСТ = ={{знак}, {рядок}}, кодуються одиниці графемного відображення (дискурси, речення, лексеми). При отриманні графемного значення тексту семантична інформація одержується з тексту повнотекстової бази даних вже із організації його знакової системи, виходячи з оформлення тексту.

На етапі морфологічного аналізу ЛП приписує кожній словоформі тексту комплекс морфологічної інформації (КМІ), який містить набір можливих альтернативних варіантів морфологічних структур (МС). У розділі розглядається математична модель, що описує

закономірності утворення зв'язків між двома словоформами в реченні, які стоять поряд. Дана модель дозволяє зняти значну частину морфологічної омонімії на етапі контекстного аналізу. $M=\{m_1, \dots, m_n\}$ – множина словоформ, n – кількість словоформ в словнику системи. Декартовий добуток $m_i \times m_j$, де $1 \le i \le n$, $1 \le j \le n$, являє собою граматичне словосполучення, знак х – означає, що між словоформами встановлені певні семантико-синтаксичні зв'язки. На множині М вводиться система предикатів S так, щоб кожній словоформі $m_i \in M$ відповідав деякий предикат $P(q_m) \in S$, рівний 1 під час підстановки комплексу морфологічної інформації, приписаного на попередньому етапі аналізу конкретній словоформі m_i і був рівний 0 в іншому випадку.

Бінарне відношення на множині словоформ речення, що стоять поряд для всіх типів граматичного підпорядкування може бути задане формулою:

$$P(q_m) \times P(q_n) = \gamma_t(q_m, q_n) \bullet P(q_m) \bullet P(q_n), \quad (10)$$

де знак х, означає операцію з'єднання КМІ словоформ і вказує на те, що дві словоформи, які стоять поряд, пов'язані між собою семантико-синтаксичним зв'язком; • — операція кон'юнкції предикатів; а множник $\gamma_t(q_m, q_n)$, $t = 1,2,3$ (узгодження, управління, примикання, відповідно) виключає із формули (10) ті МС поряд стоячих словоформ, які не узгодяться при даному типі граматичного підпорядкування.

$$\gamma_1(q_n, q_m) = q_n^{x1} q_m^{y3} \vee q_n^{x2} q_m^{y3} \vee q_n^{x1} q_m^{y4} \vee q_n^{x2} q_m^{y4} \vee q_n^{x2} q_m^{y7},$$

$$\gamma_2(q_n, q_m) = q_n^{x1} q_m^{y2} \vee q_n^{x2} q_m^{y2} \vee q_n^{x9} q_m^{y2} \vee q_n^{x5} q_m^{y2} \vee q_n^{x6} q_m^{y2}, \quad (11)$$

$$\gamma_3(q_n, q_m) = q_n^{x5} q_m^{y9} \vee q_n^{x6} q_m^{y9} \vee q_n^{x4} q_m^{y9} \vee q_n^{x1} q_m^{y9} \vee q_n^{x2} q_m^{y9} \vee q_n^{x5} q_m^{y10} \vee q_n^{x10} q_m^{y5},$$

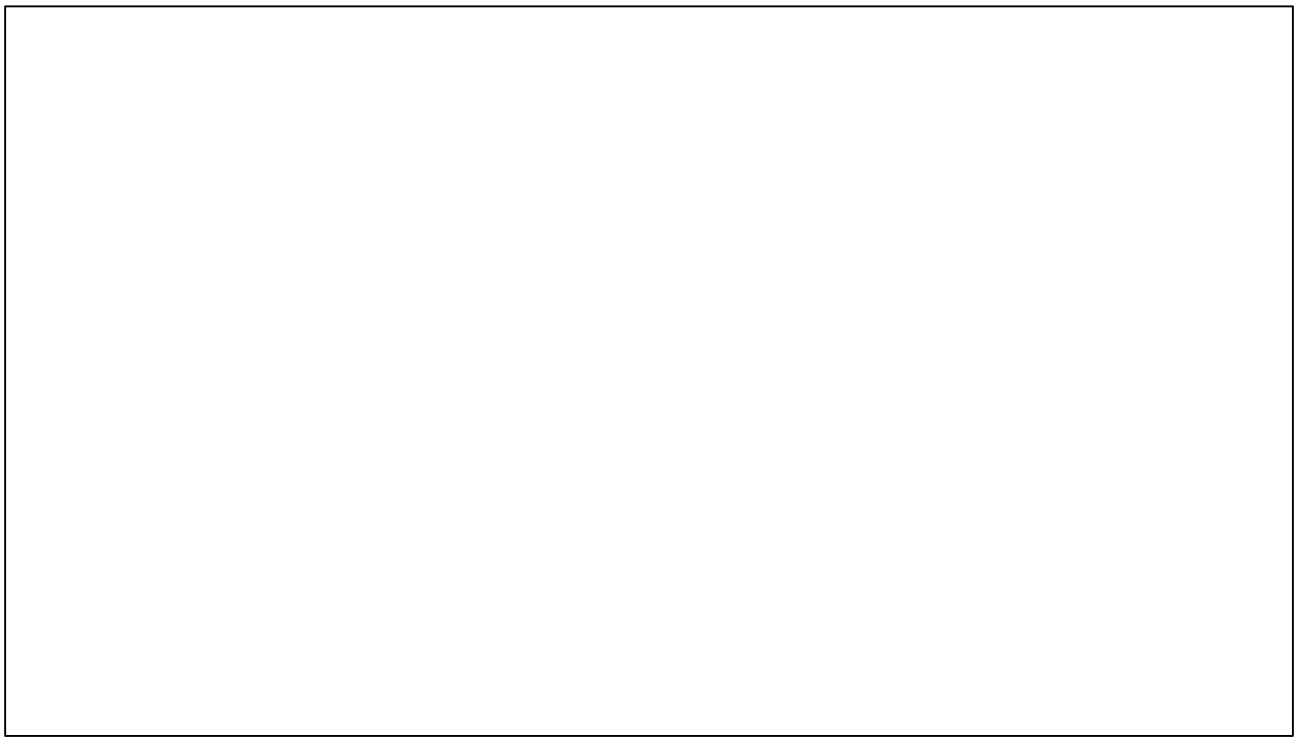
де x_i $(1 \le i \le 10)$ – *перша словоформа словосполучення;* y_j $(1 \le j \le 10)$ – *друга словоформа словосполучення;* x_1, y_1 – *іменник, називний відмінок* x_2, y_2 – *іменник, непрямий відмінок*; x_3, y_3 – *прикметник*; x_4, y_4 – *дієприкметник;* x_5, y_5 – *дієслово не минулого часу*; x_6, y_6 – дієслово *минулого часу*; x_7, y_7 – *порядковий числівник*; x_8, y_8 – *кількісний числівник;* x_9, y_9 – *прийменник* x_{10}, y_{10} – *прислівник*.

Під час підстановки КМІ першої і другої словоформи словосполучення, отримані на етапі морфологічного аналізу, в формулу (11), предикати, що описують тип словосполучення, не властивий даним словоформам, обертаються в нуль. Ті ж предикати,

які приймають значення 1, дозволяють істотно зменшити кількість можливих варіантів сполучень між словоформами.

На етапі аналітико-синтетичного аналізу використовано модель предметної області у вигляді концептуального рубрикатора або тезауруса, що дозволяє розглядати не стійкі зв'язки між мовними одиницями, а стійкі зв'язки між поняттями, які носять енциклопедичний характер. Даний підхід дає можливість створювати модель предметної області незалежно від мов і реально полегшує розширення списку мов, з якими може працювати система. В дисертації розроблена модель предметної області на прикладі понять, об'єктів і відносин з області комп'ютерних технологій і Internet, де предикат $P(x_1,x_2)$ показує відношення частина-ціле, а предикат $Q(x_1,x_2)$ відношення рід-вид, An – дескриптор (рис.3). Список дескрипторів, що ставляться до предметної області комп'ютерних технологій та Internet, і розглядаються в дисертаційному дослідженні приведено в додатку англійською, російською та українською мовами.

У четвертому розділі приведено опис програмного комплексу Texs, реалізованого по запропонованій автором моделі і дається оцінка його ефективності. Розділ містить практичні рекомендації із використання результатів дисертаційного дослідження в системах автоматичного реферування. Обгрунтовується можливість реалізації методів і моделей дисертаційного дослідження під час розробки інформаційно-пошукових засобів української частини мережі Internet.



комп'ютерних технологій і Internet (комп'ютерна архітектура).

Алгоритм роботи системи рубрикації документів за вузькими тематичними класами складається з двох частин. Перша частина являє собою етап автоматичного навчання системи. На цьому етапі відбувається автоматичне формування рубрикатора, під час якого відбувається розподіл семантичного простору даної предметної області на мікрообласті: формування формалізованого образу рубрики (ФОР). На другому етапі роботи система використовує еталони, отримані на етапі навчання для розділення текстів документів, що аналізуються, за мікротемами (тобто віднесення документів до певної класифікаційної підрубрики). Алгоритм роботи включає п'ять блоків.

1. Блок графемної обробки, що визначає мову текстів, які аналізуються і що виділяє структурно-певні фрагменти повнотекстової бази даних (заголовок, перший абзац документа, перше речення першого абзацу і т.д.).
2. Блок морфологічної обробки вводиться для обліку словозмінних і словотворчих форм мови. На цьому етапі створюється словник квазіоснов у вигляді множини, яка містить всі лексикографічні варіанти словозмінних основ ключових термінів.
3. Блок статистичної обробки формує інформаційне уявлення кожного тексту бази, яка розглядається у вигляді алфавітного словника ключових слів (АСКС), із приписаними ним ваговими коефіцієнтами.
4. Блок компараторної ідентифікації реалізовує дескрипторно-текстовий предикат *P(t, r),* заданий на декартовому добутку T x R множин текстів, які розглядаються і ключових слів. Внаслідок реалізації цього блоку отримують еталони конкретних мікротем, які представляють класи ключових термінів.
5. Етап систематизації (рубрикації) документів. На цьому етапі, внаслідок процедури графемної, морфологічної і статистичної обробки, кожному тексту, що надходить на вхід системи, приписується АСКС. Порівнюючи словники ключових слів кожного тексту з діючими еталонами, підсистема АІБС класифікує множину текстів повнотекстової бази даних за вузькими предметними рубриками.

Як показали результати реалізації, комплекс прикладних програм Texs являє собою динамічну підсистему систематизації і предметизації автоматизованої інформаційно-біблиотечної системи, яка швидко настроюється на нову семантичну область та інваріантна до мови текстів, що класифікуються. Незалежність формування інформаційно-лінгвістичного і програмного забезпечення, дозволяє створити відкриту, адаптивну, ієрархічну та модульну структуру алгоритму.

Результати дисертаційних досліджень були використані під час розробки підсистеми аналітико-синтетичної обробки текстів українською і російською мовами АІБС

в науково-методичному відділі Центральної наукової бібліотеки Харківського національного університету ім. В.Н. Каразіна. Проведені дослідження показали, що у 73% випадків рубрикація текстів повнотекстової бази, здійснених системою Texs, збігається із розподілом текстів за предметними рубриками, здійсненими експертами.

У додатках приведено фрагменти тексту програм комплексу Texs; список дескрипторів англійської, української і російської мов, що використовується в моделі предметної області комп'ютерних технологій і Internet; бібліографічний список документів, приклад рубрикації яких розглянуто в роботі; а також акти про впровадження результатів дисертаційного дослідження.

ВИСНОВКИ

1. Розроблено метод класифікації документальної інформації, який засновується на моделюванні системи асоціацій наукових і технічних понять і областей, використання цього методу дозволяє підвищити релевантність та повноту інформації АІБС, що видається.

2. Розроблені інструментальні засоби компараторної ідентифікації, необхідні для моделювання процедур класифікації, предметизації і систематизації документів: математичні засоби опису дескрипторно-текстового предиката, функцій розуміння тексту і лексичних одиниць, які виражають теми (поняття, рубрики).

3. Проведена структурно-функціональна класифікація лінгвістичного забезпечення автоматизованої інформаційно-бібліотечної системи. Розроблена схема лінгвістичного процесора, що враховує особливості інформаційних повідомлень, які надходять в АІБС. Побудовані моделі обробки текстових повідомлень на графемному, морфологічному і семантико-синтаксичному рівнях роботи ЛП.

4. Запропоновано метод моделювання інтелектуальної функції людини по розумінню і класифікації за змістовими ознаками лексичних одиниць мови. Вперше введене і обгрунтоване поняття предиката інтелектуальної аналітико-синтетичної обробки документа, який дозволяє формально подати відносини між розумінням предмета, що розглядається в тексті документа, і поняттям, що виражається відповідним йому ключовим словом. Проведена формалізація процесу дескрипторизації, яка усуває неоднозначність у вигляді омонімії і полісемії ключових слів і що дозволяє здійснити їх групування за класах умовної і безумовної еквівалентності.

5. Розроблено і програмно реалізовано алгоритм формування рубрикатора, що дозволяє автоматизувати роботу системи поділу документів повнотекстової бази даних за

вузькими підтемами, а ключових термінів — за підрубриками. Запропонований алгоритм дозволяє розробити динамічний рубрикатор АІБС, який настроюється на нові ПО і є інваріантним до мови текстів, що класифікуються.

6. Розглянуто коло найбільш важливих завдань автоматизованої обробки текстової інформації, пов'язаних із реферуванням і анотуванням текстів повнотекстової БД. Вироблені практичні рекомендації по використанню результатів дисертаційного дослідження в системах автоматичного реферування. Окреслені методи і моделі, які можуть бути використані при розробці інформаційно-пошукових засобів української частини мережі Internet.

7. Результати проведених дисертаційних досліджень були використані під час розробки підсистеми аналітико-синтетичної обробки текстів українською і російською мовами АІБС в науково-методичному відділі Центральної наукової бібліотеки Харківського національного університету ім. В.Н. Каразіна, а також під час виконання теми "Корекція-А" в Харківському військовому університеті.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Хайрова Н.Ф., Замаруева И.В. Машинный перевод: Учеб. пособие. – Х.: Око, 1998. – 82 с.

2. Хайрова Н.Ф. Компараторная идентификация документов в полнотекстовой базе данных // АСУ и приборы автоматики.– Х., 1999.– N 109.– С.67-76.

3. Хайрова Н.Ф., Шаронова Н.В., Ситников Д.Э. Моделирование аналитико-синтетической обработки каталогизатором текста документа // Вестн. Харьк. гос. политехн. ун-та. – Х., 1999. – Вып.43.– С. 82-91.

4. Ситников Д.Э., Шаронова Н.В., Хайрова Н.Ф. Моделирование семантико-синтаксических отношений грамматических словосочетаний // Пробл. бионики. –Х., 1999. –Вып.50.– С. 179-184

5. Шаронова Н.В., Хайрова Н.Ф. Направления совершенствования современных систем автоматического реферирования// Вестн. Херсон. гос. техн. ун-та.– Херсон, 1999.– N 1.– С.78-80.

6. Шаронова Н.В., Хайрова Н.Ф. Построение модели базы знаний в автоматизированной информационно-библиотечной системе// Вестн. Херсон. гос. техн. ун-та. – Херсон, 1998.– N2 .–С.105-110

7. Хайрова Н.Ф. Современные аспекты автоматического реферирования// Ученые записки Харьковского гуманитарного института "Народная украинская академия.– Х.: Око, 1997.– Т.3– С.364-373.

8. Хайрова Н.Ф. Преподавание основных методов семантической обработки текстов в рамках курса "Машинный перевод" // Концепция приватного образования: принципы, содержание, проблемы реализации / М-во образования Украины; ХГИ “НУА”. – Х., 1998.– С.98-106.– Деп. В ХГТБ Украины 13.07.98, N 316 – Ук98.

9. Хайрова Н.Ф. Грамматический анализ текста на естественном языке, как важнейший этап получения знаний из этого текста// Актуальні питання впровадження інформаційних технологій у документально-комунікаційній сфері: Прогр. та матеріали III міжнар. семінару (11-13 верес. 1996 р., Харків/ Асоц. сучас. інформ.-бібл. технологій та ін.– Х., 1996.– С.73-74.

10. Хайрова Н.Ф. Требования, предъявляемые к современным автоматизированным информационно-библиотечным системам // Актуальные проблемы гуманитарных наук и их информационное обеспечение: Материалы IV Междунар. студ. науч. конф., Харьков, 26 апр. 1997.– Х., 1997.–С.38-39.

11. Хайрова Н.Ф. Состояние программного обеспечения компьютерного перевода в общей системе подготовки референтов-переводчиков// Досвід і проблеми реалізації ступеневої системи підготовки фахівців: Зб. матеріалів III міжнар. наук.-метод. конф. , Суми, 8-11 верес. 1997 р. –Суми, 1997.– С.264.

12. Шаронова Н.В., Хайрова Н.Ф. Системы автоматического реферирования текстов// Создание, интеграция, использование информационных ресурсов инновационного развития: Тез. докл. и сообщ. междунар. конф. Киев, 18-19 дек. 1997 г.– К., 1997.– С.173-175.

13. Шаронова Н.В., Хайрова Н.Ф., Ситников Д.Э. Логико-алгебраическая модель автоматизированной классификации электронных документов// Теория и техника передачи, приема и обработки информации. Телекоммуникации. Радиотехника. Электроника: Сб.науч.тр. – Х., 1999.– С.451-453.

АНОТАЦІЯ

Хайрова Н.Ф. Розробка математичного і лінгвістичного забезпечення автоматизованих інформаційно-бібліотечних систем. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – автоматизовані системи управління та прогресивні ін-формаційні

технології. – Харківський державний технічний університет радіо-електроніки, Харків, 2000.

Дисертація присвячена питанням розробки інформаційних технологій, які базуються на моделях та алгоритмах семантичної обробки документів автоматизованих інформаційно-бібліотечних систем (АІБС). У роботі вперше використаний метод компараторної ідентифікації для моделювання процедур аналітико-синтетичної обробки текстів документів. Введені дескрипторно-текстовий предикат; функції розуміння тексту і ключових термінів, які дозволили розробити алгоритми процесів систематизації і предметизації. Побудовані моделі обробки текстових повідомлень на графемному, морфологічному і семантико-синтаксичному рівнях роботи лінгвістичного процесора інформаційної системи. Створений програмний комплекс Texs, що являє собою підсистему систематизації і предметизації АІБС. Подана інформація про практичну реалізацію і ефективність розроблених методів і алгоритмів.

Ключові слова: інтелектуальна система, метод компараторної ідентифікації, семантичний аналіз, класифікація, інформаційні технології, автоматизовані бібліотечні системи.

SUMMARY

Khayrova N.F. Working out mathematical and linguistic maintenance of automated information-library systems. – Manuscript.

Thesis for a candidate's degree by speciality 05.13.06 – automated control systems and progressive information technologies. – Kharkov state technical university of radioelectronics, Kharkov, 2000.

The present thesis is on working out models, algorithms and information technology of semantic processing documents of automate information-library systems (AILS). The method of comparative identification for modeling procedures of analytic-synthetic is processing the texts of documents. The notion of descriptive-text predicate is introduced as well as the notion of text and key terms understanding function, allowing to work out algorithms of systematization and objectization procedures. Models of processing text reports on the level of graphemes as well as on morphological semantic-syntactic levels of work of a linguistic processor of an information system are built. A program complex Texs representing a subsystem for systematization and objectization of AILS is created. Information about practical realization and efficiency of worked out models and algorithms is given.

Key words: intellectual system, comparative identification method, semantic analysis, classification, information technology, automate library systems.

АННОТАЦИЯ

Хайрова Н.Ф. Разработка математического и лингвистического обеспечения автоматизированных информационно-библиотечных систем. – Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 – автоматизированные системы управления и новые информационные технологии. – Харьковский государственный технический университет радиоэлектроники, Харьков, 2000.

Диссертация посвящена вопросам разработки информационных технологий основанных на моделях и алгоритмах семантической обработки документов автоматизированных информационно-библиотечных систем (АИБС). Предлагаемые в работе разработки математического и лингвистического обеспечения АИБС основаны на формализации и моделировании интеллекта, проводящего аналитико-синтетическую обработку текстов и тесно связанную с ней предметно-тема-тическую ассоциацию в процессе организации тематического поиска. Для достижения данной цели предлагается формализовать задачу повышения релевантности и полноты выдаваемой библиотечной системой информации за счет разработки методики классификации, основывающейся на моделировании системы ассоциаций научных и технических понятий.

В работе используются модели и методы новых информационных технологий, базирующиеся на результатах, полученных при решении проблем искусственного интеллекта. Для организации узко-тематического поиска в информационных системах, работающих с полнотекстовыми базами данных, предлагается учитывать не только логическую связь терминов и понятий, но и аспекты естественного языка. Проведена структурно-функциональная классификация лингвистического обеспечения автоматизированной информационно-библиотечной системы. Разработана схема лингвистического процессора (ЛП), учитывающего особенности информационных сообщений, поступающих в АИБС.

В лингвистическом процессоре используются модели обработки текстовых сообщений на графемном, морфологическом и семантико-синтаксическом уровнях языка. Для повышения полноты и релевантности тематического поиска в работе используется дескрипторный язык с линейной (позиционно-скобочной) грамматикой, отражающей взаимное расположение лексических единиц и структурных фрагментов текста, определяемых на этапе графемного анализа.

На этапе контекстного анализа снимается часть языковой неоднозначности. Предлагаемый метод основан на математической модели, представляющей собой задачу

выявления и математического описания закономерности образования связей между двумя рядом стоящими словоформами в предложении. Из бинарного семантико-синтаксического отношения, заданного на множестве рядом стоящих словоформ предложения, исключаются те морфологические структуры, которые не согласуются при данном типе грамматического подчинения .

Этап семантической обработки текста документа представлен рубрикатором, приписывающим входным единицам информации обозначения, адекватно отражающие их содержание. Для формализации процедур аналитико-синтетической обработки текстов документов, циркулирующих в АИБС, впервые использован метод компараторной идентификации. Использование данного метода позволило моделировать функции интеллекта по систематизации и предметизации документов библиотеки. Переходя от предиката интеллектуальной аналитико-синтетической обработки текста к дескрипторно-текстовому предикату, осуществляется переход от субъективного восприятия понятий и денотатов к объективному соответствию между текстом и ключевыми терминами. Введение дескрипторно-текстового предиката; функции понимания текста и функции понимания ключевых терминов, позволило разработать математическое обеспечение процедур каталогизации, систематизации и предметизации, непосредственно связанных с организацией тематического библиотечного поиска.

В работе приведено описание практической реализации положений диссертации в программном комплексе "Texs", разработанном автором по собственной модели, и представляющем собой динамичную, быстро настраиваемую на новую семантическую область и инвариантную к языку классифицируемых текстов подсистему предметизации и систематизации АИБС. Алгоритм работы системы Texs состоит из двух частей. Первая часть представляет собой этап обучения системы, на котором происходит автоматическое формирование рубрикатора, разбивающего семантического пространства данной предметной области на микро области: формирование формализованного образа рубрики. На втором этапе работы система использует эталоны, полученные на этапе обучения, для разделения анализируемых текстов документов полнотекстовой базы данных по микро темам (т.е. отнесение документов к определенной классификационной подрубрике). В работе дана оценка эффективности программного комплекса Texs. Практические результаты работы показали, что в коэффициент релевантности текстовой информации, выдаваемой системой, близок к 0,7.

Ключевые слова: интеллектуальная система, метод компараторной идентификации, семантический анализ, информационная технология, автоматизированные библиотечные системы.