

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

Сайед Мохаммад Таухід Сіддікі



УДК [004.415:005.92]:342.5(043.5)

**МОДЕЛІ І ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ІНТЕЛЕКТУАЛЬНОЇ
ОБРОБКИ ДАНИХ ПРИ СТВОРЕННІ ПАКЕТІВ ДІЛОВОЇ
ДОКУМЕНТАЦІЇ**

05.13.06 – інформаційні технології

Автореферат

дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2009

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
Четвериков Григорій Григорович,
Харківський національний університет
радіоелектроніки, м. Харків,
професор кафедри програмного забезпечення ЕОМ

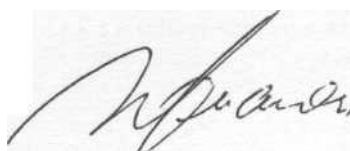
Захист відбудеться « 14 » травня 2009 р. о 14.30 годині на засіданні спеціалізованої вченої ради Д 64.050.07 в Національному технічному університеті «Харківський політехнічний інститут» за адресою: 61002, м. Харків, вул. Фрунзе, 21.

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

Автореферат розісланий «10» квітня 2009 р.

Вчений секретар

Спеціалізованої вченої ради



І.П. Гамаюн

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Єдині вимоги до оформлення та обробки документів дають змогу швидко отримати потрібну інформацію. Використання комп'ютерних засобів для оформлення та зберігання документів відкриває нові можливості для прискорення процесу обробки інформації і, як наслідок, прийняття правильного рішення й ефективного управління.

Використання комп'ютерів для оформлення, зберігання й передачі документів призвело до появи електронних документів, більша частина яких представлена у вигляді текстової інформації. Програмна обробка текстової інформації (перш за все в інформаційно-пошукових, експертних системах різних галузей застосування, у бібліографічних системах, у системах машинного перекладу, при обробці баз даних тощо) в умовах неперервного збільшення потоку науково-технічної інформації набуває все більшого значення. Ефективне використання комп'ютерної техніки в цьому напрямку неможливе без створення діючих моделей природних мов.

Організація ефективної роботи з документами, удосконалення всіх інформаційно-документальних процесів як у масштабі суспільства в цілому, так і на рівні окремих підприємств, закладів, установ є найважливішим напрямком удосконалення управлінської діяльності. Рациональна організація функціонування систем документального забезпечення управління є однією з найважливіших ланок зміцнення апарату управління й суттєвим елементом його стабільності в галузі прийняття ефективних рішень.

Комп'ютерне моделювання процесів обробки текстів різної складності дасть змогу автоматизувати різні види інтелектуальної діяльності людини, розширити його можливості щодо прийняття ефективних управлінських рішень. Ефективність автоматизованих інформаційних систем (АІС) визначається їхньою здібністю оброблювати інформацію неформалізовану або слабо формалізовану. Задача створення пакетів ділової документації є однією з таких актуальних галузей обробки інформації.

У межах окресленої проблеми важливим є наукове завдання розробки методів, моделей, алгоритмів і програм, які здійснюють моделювання процесів інтелектуальної обробки даних при створенні пакетів ділової документації з метою визначення їх основних характеристик для побудови інформаційного, математичного, лінгвістичного і програмного забезпечення систем електронного документообігу. Значний внесок у розвиток інформаційних технологій, пов'язаних з автоматизацією організаційного управління, зробили такі видатні вчені, як Г.Г. Асєєв, В.М. Брябрін, В.М. Глушков, І.П. Кузнецов, Д.О. Поспєлов, С.В. Назаров, В.ІІІ. Рубашкін та ін.

Усе вищезазначене обумовлює **актуальність** обраного напрямку дисертаційного дослідження, яке пов'язане з розвитком методів автоматизації процесів створення пакетів ділової документації із застосуванням сучасних методів інтелектуальної обробки текстової інформації.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана відповідно до напрямків наукових досліджень кафедри інтелектуальних комп'ютерних систем НТУ «ХПІ» у межах держбюджетної теми «Розробка математичних моделей і методів розв'язання задач інтелектуальної обробки інформації» (№ ДР 0108U003926), у якій здобувач брав участь як виконавець.

Мета і задачі дослідження. Метою дисертаційної роботи є підвищення ефективності та якості створення пакетів ділової документації на основі розробки математичних моделей та інформаційної технології інтелектуальної обробки даних. Відповідно до зазначеної мети у роботі поставлено та розв'язано такі задачі:

- 1) провести системний аналіз основних напрямків автоматизації діловодства і створення пакетів ділової документації у системах електронного документообігу як об'єктів автоматизованої обробки даних;
- 2) провести аналіз загальних проблем створення електронного архіву документів, дослідити основні функції електронного архіву документів;
- 3) розробити математичну модель інтелектуальної обробки даних, яка дозволяє перейти до інтелектуально-орієнтованих методів розробки документів;
- 4) удосконалити методи інтелектуальної обробки даних на основі застосування теорії категорій і логічних мереж;
- 5) розробити модель логічної мережі для певних типів документів, які входять до складу пакету документації з тендерних угод;
- 6) розробити програмний комплекс розв'язання рівнянь алгебри предикатів для реалізації отриманих логічних мереж;
- 7) дослідити структуру й компоненти системи електронного документообігу, яка дозволяє моделювати створення та обробку документів у реальному часі.

Об'єктом дослідження є процеси створення пакетів ділової документації й документообігу в системі підготовки міжнародних контрактів.

Предметом дослідження є побудова інтелектуальних моделей обробки даних при формуванні документів.

Методи дослідження засновані на комплексному використанні теорії інтелекту, теорії категорій та логічних мереж, апарату алгебри скінченних предикатів для моделювання семантичної структури документів і документообігу. Основи теорії множин та теорії відношень використовуються для переходу до алгебро-логічних моделей. Імітаційне комп'ютерне моделювання – для практичної реалізації одержаних результатів.

Наукова новизна одержаних результатів визначається наступним:

1. Дістали подальший розвиток:

- комплексний підхід до моделювання процесів діловодства з урахуванням особливостей та типів документів, який забезпечує можливість формалізувати й автоматизувати процес створення пакету ділової документації;

- метод аналізу й синтезу документів, заснований на застосуванні теорії категорій і логічних мереж, який дозволяє сформувати документи тендерних угод з урахуванням їх змістовного наповнення, що забезпечує підвищення якості документів та швидкості їхнього виконання.

2. *Уперше* розроблено алгебро-логічну модель процесу діловодства, яка відрізняється від відомих моделей використанням методу компараторної ідентифікації, теорії модифікованих категорій і апарату логічних мереж. Модель дозволяє відображати відношення між інформаційними параметрами документів і процесами їхньої обробки, що забезпечує підвищення точності моделювання змісту документів.

3. *Удосконалено* математичну модель аналітико-синтетичної обробки текстової інформації в документах і створено інформаційну технологію інтелектуальної обробки даних при створенні пакетів ділової документації, що підвищує рівень автоматизації обробки текстової інформації, яка міститься в документах, та дозволяє підвищити якість створених документів.

Практичне значення одержаних результатів. Розроблені в дисертації математичні засоби, моделі, алгоритми й програмні системи призначені для створення логічних мереж, які є основою для обчислювальних модулів паралельної дії. Математичні результати роботи можуть бути використані в системах автоматичної обробки природної мови (розробка природно-мовних інтерфейсів, системи машинного перекладу тощо). Логічні мережі можуть бути ефективні в системах автоматизованої розробки пакетів ділової документації, у системах електронного документообігу, а також для моделювання процесів діловодства в інтерактивному режимі.

Результати дисертаційного дослідження знайшли практичне застосування в ЗАТ «Індустріальна науково-технологічна компанія ІНТЕК» (м. Харків), ТОВ «Хартеп» (м. Харків) у вигляді алгебро-логічних моделей, які були використані при розробці системи електронного документообігу. 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».

Програмне забезпечення, розроблене в дисертації, використовується при виконанні курсових і дипломних робіт на кафедрах АСУ й інтелектуальних комп'ютерних систем НТУ «ХПІ».

Особистий внесок здобувача в розробку наукових результатів. Всі основні результати дисертаційної роботи, що виносяться на захист, отримані здобувачем особисто, серед них: підхід до використання методів Data Mining і Text Mining для обробки текстової інформації в інформаційних системах; математична модель ідентифікації документів у системах електронного

документообігу; спосіб ефективного аналізу текстової інформації за допомогою методу компараторної ідентифікації та ситуаційно-текстового предикату, процедура використання предикатних категорій та логічних мереж для представлення інформації у документах.

Апробація результатів дисертації. Результати дисертаційної роботи доповідалися та обговорювалися на: Міжнародній конференції «Інформаційні технології в освіті і управлінні» (Нова Каховка, 2006, 2007); Міжнародній конференції «MegaLing'2006. Горизонти прикладної лінгвістики та лінгвістичних технологій» (Крим, с. Партеніт, 2006 р.); XIV Міжнародній науково-практичній конференції «Інформаційні технології: наука, техніка, технологія, освіта, здоров'я» (м. Харків, 2006 р.); 12 Міжнародному молодіжному форумі «Радіoeлектроніка й молодь в 21 столітті» (м. Харків, 2008 р.), на семінарах кафедри інтелектуальних комп'ютерних систем НТУ «ХП».

Публікації. Результати дисертації опубліковані в 4 наукових працях, серед яких 3 статті у фахових виданнях ВАК України.

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел, додатків. Повний обсяг дисертації становить 167 сторінок, 22 рисунка по тексту, 3 додатки на 18 сторінках, 150 найменувань використаних літературних джерел на 14 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність теми роботи з точки зору перспектив розвитку та удосконалення систем автоматизованої обробки текстової інформації при формуванні різних типів документів. 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, наведена структура дисертації.

Перший розділ присвячений аналізу досягнень у галузі автоматизації діловодства та документообігу, використання інформаційних технологій у процесах діловодства та документообігу на сучасному етапі. Проведено аналіз основних проблем створення електронних архівів документів та розгляд основних їх функцій. Використання комп'ютерних засобів для оформлення та зберігання документів відкриває нові можливості для прискорення процесу обробки інформації і, як наслідок, прийняття рішення й ефективного управління. На основі проведеного аналізу літературних джерел виділено основні характерні задачі автоматизації документообігу, засновані на формалізації їх функцій, урахуванні різних типів документів, що оброблюються, і типів користувачів системи. Проаналізовано основні функції електронних архівів

документів, які пов'язані з обробкою текстової інформації. Пошуковий механізм повинен мати певний інтелект, тобто забезпечувати пошук близьких граматичних конструкцій, а також близьких за змістом слів. Задачі пошуку за семантичними ознаками є найбільш складними й актуальними для систем, у яких важливе місце посідає семантична інформація. Ще більш складними функціями є автоматичне розпізнавання змісту образу документа й формування документа, який містить його текст, та розпізнавання змісту форм. Розглянуто роботизовані бібліотеки як системи довгострокового зберігання інформації та основні задачі обробки інформації в них. Обґрунтовано необхідність розвитку математичного апарату моделювання інтелектуальних функцій людини для формалізації обробки даних при створенні ділової документації. Зроблено огляд основних напрямків інтелектуальної обробки даних.

На підставі критичного аналізу існуючих невирішених проблем та задач автоматизованого формування документації сформульовані й обґрунтовані цільові задачі наукових досліджень, які були виконані в дисертаційній роботі.

Другий розділ присвячено розробці математичного апарату для моделювання процесів обробки даних при створенні електронного документа. Визначено математичний інструментарій для представлення інформації в документах на основі теорії категорій, логічних мереж та алгебри предикатів.

Безоб'єктна класична категорія є одним із видів алгебр і задається множиною M , елементи якої називаються *морфізмами*, та єдиною частковою бінарною операцією fg множення морфізмів, яка відображає декартів добуток $M \times M$ у M . Оскільки операція множення є у загальному випадку частковою, існує єдиний лівий одиничний морфізм e_f і єдиний правий одиничний морфізм $e_{f'}$ для будь-якого морфізма $f \in M$.

У роботі в об'єктній категорії додатково до морфізмів вводиться поняття об'єктів. Множина об'єктів категорії K позначається $Ob K$. Якщо $A \in Ob K$, то A є об'єктом категорії K або K – об'єктом. Для кожного морфізма $f \in Mor K$ існує єдина пара об'єктів A і B .

Класична категорія припускає різні інтерпретації, у тому числі проєктивну й предикатну інтерпретації. Предикатна категорія $Pred$ задається на деякому універсумі U . У ролі об'єктів $A, B, C, \dots$ категорії $Pred$ використовуються довільні підмножини універсуму U . У ролі множини об'єктів $Ob Pred$ вибираємо систему всіх підмножин універсуму U . У ролі морфізма $f : A \rightarrow B$ категорії $Pred$ використовується довільний лінійний логічний оператор $F_j(P) = Q$, який перетворює предикат P у предикат Q .

Подібні лінійні логічні оператори можуть бути використані для збереження знань про правила утворення морфемних ланцюжків, словосполучень, речень, а також для отримання

семантики словоформ за відомими семантичними значеннями морфем, семантики словосполучень за відомими семантичними значеннями окремих слів тощо.

Розглянуто питання предикатної інтерпретації добутку морфізмів.

Розглянуто можливість використання предикатних категорій для формалізації знань, які представлені за допомогою логічних моделей. В основі логічних моделей представлення знань полягає формальна система, яка представлена четвіркою виду $M = \langle T, P, A, B \rangle$. Логічну систему представлення текстів можна подати у вигляді генератора текстів, у якому з вихідної множини A одержують нові тексти, які складають множину текстів (знань), які виводяться $M_{\text{вз}}$.

Теорія категорій і її предикатна інтерпретація дозволяють описувати процеси формування множини знань у базах знань логічного типу. Використовуючи предикатні категорії для описання процесу формування баз знань, множину правил виводу можна зберігати у вигляді ядер лінійних операторів, а сам механізм формування знань – у вигляді лінійних операторів, які представлені за допомогою формул алгебри предикатів. Схематична реалізація лінійних операторів дозволяє створити процесор обробки й формування знань, який включає базу знань і блок логічного виводу на знаннях. Теорія категорій надає можливості чітко і наочно описувати процеси формування й обробки знань у вигляді категоричних діаграм і є реальною основою створення систем інтелектуальної обробки текстової інформації.

У **третьому розділі** розглядаються основні задачі моделювання інтелектуальної обробки даних при створенні електронного документа. Зроблено порівняльний аналіз обробки текстової інформації за допомогою технологій Text Mining та методу компараторної ідентифікації лінгвістичних об'єктів. Технологія глибинного аналізу тексту – Text Mining – це інструментарій, який дозволяє аналізувати великі обсяги інформації у пошуку тенденцій, шаблонів та взаємозв'язків, які здібні допомогти у прийнятті стратегічних рішень. Крім того, Text Mining – це новий вид пошуку, який, на відміну від традиційних підходів, не тільки знаходить списки документів, формально релевантні запитам, але й забезпечує достатньо високий рівень аналізу з метою прийняття ефективного рішення.

Об'єкти, які оброблюються інформаційними системами, є дискретними, скінченними й детермінованими, що дозволяє використовувати при обробці цих об'єктів метод компараторної ідентифікації.

У цьому ж розділі представлено розробку алгебраїчної моделі ідентифікації документів типу «лист-запит», які за різними видами продуктів («хімічний» (органічний і неорганічний), «харчовий», «продукт/вироби» «техніка/обладнання») мають схожу структуру, але й принципіальні відзнаки.

Виділено формальні ознаки документа типу «листи-запити», які є однаковими для усіх типів документів, такі як: D – документ-контракт; X_1 – предмет контракту; X_2 – ціна товару та обсяг контракту; X_3 – термін постачання і специфікації товару; X_4 – умови постачання й прийняття товару; X_5 – термін і умови платні; X_6 – повідомлення «продавцем/постачальником»; X_7 – страхування товару; X_8 – гарантії якості; X_9 – штрафні санкції; X_{10} – упаковка й маркування; X_{11} – форс-мажор; X_{12} – арбітраж; X_{13} – інші умови; X_{14} – процедура; X_{15} – bank information .

Значення змінних ознак і предметні змінні: Y_1 – фіто-санітарія; Y_2 – радіоактивність; Y_3 – структура товару, яка впливає на упаковку товару; Y_4 – клас безпеки при перевозці; Y_5 – упаковка; Y_6 – продукт, придатний для вживання людиною (Not Genetically modified organism (NON-GMO)).

Розглянуто основні типи документів і побудовано структуру логічної мережі документа за типом товару.

У результаті одержані наступні рівняння для визначення (ідентифікації) документів за типом товару.

Описано кожну з ознак логічної мережі.

Четвертий розділ присвячено практичній реалізації одержаних результатів дисертаційного дослідження. У цьому ж розділі наведено вимоги до баз даних при формуванні тексту документа, а також викладено питання моделювання структури бази даних предметної області.

Обґрунтовано вибір такої структури бази даних, яка забезпечує узгоджену взаємодію усіх її компонентів відповідно до заданих функціональних вимог у межах заданих обмежень. У роботі надано перевагу СУБД MS SQL Server, оскільки подана СУБД спроектована для роботи у багато-користувальницькому режимі; використовує надійну систему збереження даних; підтримує мову структурованих запитів Transact SQL. У якості програмного середовища реалізації було обрано середовище Builder 6 фірми Borland.

У розділі розглянуто особливості моделювання процесів створення пакетів ділової документації у ЗАТ «Індустріальна науково-технологічна компанія ІНТЕК». Запропоновані у дисертаційному дослідженні моделі були використані при розробці інформаційного, математичного, алгоритмічного й програмного забезпечення системи економічного документування при оформленні документації з ЗАТ «Індустріальна науково-технологічна компанія ІНТЕК». У результаті аналізу множини документів (додаток В), розроблено алгебро-логічні моделі типових документів.

Структура розробленого програмного комплексу AIB (Assistant in Business) складається з двох рівнів: інтерфейсу користувача й робочого вікна, яке, у свою чергу, складається з двох вікон:

дерева-документа й вікна відображення змісту документа. Діалог з користувачем реалізовано з максимально можливим використанням миші для вибору необхідної дії з системи меню, мінімізована необхідність введення даних з клавіатури. При розробці програми було використано архітектуру управління подіями та об'єктно-орієнтований підхід, при якому програма представляється у вигляді сукупності об'єктів. В основі формування дерева документа програми лежать моделі, методи й алгоритми, розроблені в дисертаційному дослідженні. Інтерфейс програми АІВ дозволяє навіть недосвідченому виконавцю формувати необхідні розділи документів для певного виду продукції.

Запропоновані в дисертаційному дослідженні моделі були використані при розробці інформаційного, математичного, алгоритмічного і програмного забезпечення СЕД в ТОВ «Хартеп», а також при підготовці пакетів документів фірми «Consultants & Engineers Limited» (м. Харків).

На рис. 7 показано роботу програми на матеріалі ТОВ «Хартеп». Програмне забезпечення, розроблене в роботі, забезпечує користувачеві дружній інтерфейс, із зручним меню й системою допомоги й довідкою.

Для створення програмної системи було використано мову програмування C++, яка поєднує процедури й елементи об'єктно-орієнтованих й компонентно-орієнтованих мов програмування. Середовище розробки C++ включає інструменти для візуального конструювання користувацького інтерфейсу з основними перевагами – швидкою розробкою прототипів програми, розробкою додатків баз даних й компонентним способом створення програм, які працюють під управлінням операційної системи Windows.

На базі розробленої програми, враховуючи, що методики розрахунку ефекту від впровадження зазвичай є унікальними для кожної організації, яка проводить подібну оцінку, у роботі розглянуто певні практичні методики й підходи, які були використані для розрахунку ефекту від впровадження систем електронного документообігу у фірмі «Consultants & Engineers Limited».

Таким чином, обґрунтовано доцільність застосування запропонованих методів інтелектуальної обробки інформації для удосконалення систем автоматизованого формування пакетів ділової документації.

ВИСНОВКИ

Основний науково-практичний результат роботи полягає в розробці й обґрунтуванні методики інтелектуальної обробки даних, заснованої на використанні теорії категорій та логічних мереж, при автоматизованому створенні пакетів ділової документації та її застосуванні для

підвищення якості документації, що забезпечує поліпшення якості прийняття управлінських рішень.

У процесі виконання роботи отримані наступні результати:

1. Проведено системний аналіз напрямків автоматизації діловодства і створення пакетів ділової документації в системах електронного документообігу. Зроблено висновок, що головною проблемою в цій галузі є автоматизоване створення пакетів ділової документації з використанням інтелектуальних методів обробки даних. Проведено аналіз проблем створення електронного архіву документів, досліджено його основні функції.

2. Проведено порівняльний аналіз обробки текстової інформації за допомогою технологій Text Mining і методу компараторної ідентифікації. Показано, що об'єкти, які оброблюються інформаційними системами (документи), є дискретними, скінченними й детермінованими, що дозволяє використовувати при обробці цих об'єктів метод компараторної ідентифікації.

3. У якості математичного апарату для моделювання процесів обробки текстових даних вибрано теорію категорій. Проведено аналіз властивостей категорій й можливостей предикатних категорій для представлення текстової інформації. Дослідження теорії категорій дозволило узагальнити поняття безоб'єктної категорії. Застосовано універсальний математичний апарат алгебри предикатів, введено інтерпретацію категорії в термінах алгебри предикатів – предикатна категорія, розглянуто обидва випадки: категорії з об'єктами та безоб'єктної категорії.

4. Показано застосування предикатних категорій для формалізації знань, які представлено за допомогою логічних моделей. Використано поняття модифікованої категорії. З урахуванням вимог до поняття модифікованої категорії узагальнено предикатну інтерпретацію категорії.

5. Розроблено математичну модель інтелектуальної обробки даних, яка дозволила перейти до інтелектуально-орієнтованих методів розробки документів. Запропоновано алгебраїчну модель ідентифікації документів типу «лист-запит». Виділено формальні ознаки документа типу «лист-запит», визначено області їх змінювання. Розглянуто основні типи документів, які входять до складу міжнародних контрактів, і побудовано структуру логічної мережі пакету документації.

6. Сформульовано вимоги до баз даних при формуванні тексту документа. Виділено основні параметри таким чином, що досягається компроміс між функціональними, інформаційними, апаратними, архітектурними й технологічними вимогами до баз даних.

7. Розроблено програмний комплекс AIB (Assistant in Business), який на основі реалізації логічних мереж здійснює функції поетапного формування документа в діалоговому режимі з узгодженням вибору та наповнення необхідних рубрик документа.

8. Розглянуто практичні методики й підходи, які були використані для підрахунку ефекту від упровадження систем електронного документообігу у фірмі «Consultants & Engineers Limited».

Виділено і проаналізовано кількісні та якісні показники підрахунку ефекту від упровадження моделей і алгоритмів.

9. Результати роботи впроваджено при розробці інформаційного, математичного, алгоритмічного й програмного забезпечення при оформленні документації в ЗАТ «Індустріальна науково-технологічна компанія ІНТЕК» (м. Харків), у ТОВ «Хартен» (м. Харків), а також використані при підготовці пакетів документів фірми «Consultants & Engineers Limited» та в навчальному процесі на кафедрах АСУ та інтелектуальних комп'ютерних систем при викладанні спеціальних дисциплін «Інформаційно-ресурсне забезпечення лінгвістичної діяльності», «Комп'ютерні системи перекладу й документування інформації».

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

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

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

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

Здобувачем запропоновано математичну модель ідентифікації документів у системах електронного документообігу.

3. Сайед Мохаммад Таухид Сиддики. Идентификация документов в системах электронного документооборота / О. В. Канищева, Сайед Мохаммад Таухид Сиддики // Вестн. Херсонского нац. техн. ун-та. – Херсон: ХНТУ, 2007. – № 4(27). – С. 273-276.

Здобувачем на основі аналізу існуючих моделей запропоновано спосіб ефективного аналізу текстової інформації за допомогою методу компараторної ідентифікації та ситуаційно-текстового предикату.

4. Сайед Мохаммад Таухид Сиддики. Обработка текстовой информации с помощью технологий Text Mining и компараторной идентификации / Н. В. Шаронова, Сайед Мохаммад Таухид Сиддики, О. В. Канищева // MegaLing'2006 Горизонти прикладної лінгвістики та лінгвістичних технологій // Доповіді міжнародної конференції. 20-27 вересня 2006, Україна, Крим,

Партеніт / Укр. мовно-інформаційний фонд НАН України, Таврійський національний університет ім. В.І. Вернадського. – Сімферополь: Вид-во "ДиАйПи", 2006. – С. 231-232.

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

АНОТАЦІЇ

Сайед Мохаммад Таухід Сіддікі. *Моделі і інформаційна технологія інтелектуальної обробки даних при створенні пакетів ділової документації. – Рукопис.*

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Національний технічний університет «Харківський політехнічний інститут», Харків, 2009.

Дисертація присвячена розробці моделей та інформаційної технології створення пакетів ділової документації, яка базується на методах інтелектуальної обробки даних. У якості математичного апарату для моделювання процесів обробки текстових даних у документах вибрана теорія категорій. Проведений аналіз властивостей категорій та можливості предикатних категорій для представлення текстової інформації.

Уведена інтерпретація категорії в термінах алгебри предикатів – предикатна категорія, розглянуті обидва випадки: категорія з об'єктами та безоб'єктна категорія. Показано, що ядро лінійного логічного оператора можна розглядати як знання або правила отримання знань, а сам лінійний логічний оператор як механізм здійснення запиту для отримання нових знань. Показано застосування предикатних категорій для формалізації даних, які представлені за допомогою логічних моделей. Досліджені властивості добутку морфізмів. Функціонування логічної моделі представлення знань відтворено у вигляді послідовності відображень та категоріальних діаграм.

Удосконалено засоби інтелектуальної обробки даних, які засновані на застосуванні теорії категорій та логічних мереж. Запропонована алгебраїчна модель ідентифікації документів типу «лист-запит». Виділені формальні ознаки документа типу «лист-запит», визначені області їх змінення. Розглянуті основні типи документів та побудована структура логічної мережі документа за типом товару. Розроблено програмний комплекс розв'язання рівнянь алгебри предикатів для реалізації отриманих логічних мереж.

Ключові слова: автоматизована переробка інформації, електронний документообіг, теорія категорій, компараторна ідентифікація, інтелектуальна обробка даних, логічні мережі.

Сайед Мохаммад Таухид Сиддики. *Моделі и информационная технология интеллектуальной обработки данных при создании пакетов деловой документации. – Рукопись.*

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 – информационные технологии. – Национальный технический университет «Харьковский политехнический институт», Харьков, 2009.

Диссертация посвящена разработкам моделей и информационной технологии интеллектуальной обработки данных при создании пакетов деловой документации. В исследовании проведен анализ основных направлений автоматизации делопроизводства и создания пакетов деловой документации в системах электронного документооборота, проблем создания электронного архива документов и его основных функций, разработана математическая модель интеллектуальной обработки данных, позволяющая перейти к интеллектуально-ориентированным методам обработки документов. В качестве математического аппарата для моделирования процессов обработки текстовых данных в документах выбрана теория категорий, аппарат логических сетей и метод компараторной идентификации. Проведен сравнительный анализ обработки текстовой информации с помощью технологий Text Mining и компараторной идентификации. Изучены свойства категорий и возможности предикатных категорий для представления текстовой информации.

Введена интерпретация категории в терминах алгебры предикатов – предикатная категория, рассмотрены оба случая: категории с объектами и безобъектной категории. Показано, что ядро линейного логического оператора можно рассматривать как знания или правила получения знаний, а сам линейный логический оператор как механизм выполнения запроса для получения новых знаний. Показано применение предикатных категорий для формализации знаний, представленных с помощью логических моделей. Исследованы свойства произведения морфизмов. Функционирование логической модели представления знаний воспроизведено в виде последовательности отображений и категорных диаграмм. Использовано понятие модифицированной категории. Обобщено понятие предикатной интерпретации категории.

Усовершенствована интеллектуальная обработка данных на основе применения теории категорий и логических сетей. Предложена алгебраическая модель идентификации документов типа «письма-заявки». Выделены формальные признаки документа типа «письма-заявки», определены области их изменения. Рассмотрены основные типы документов и построена структура логической сети документа по типу товара.

Разработана модель логической сети для некоторых типов документов, входящих в состав пакета документации по тендерным соглашениям. Разработан программный комплекс решения уравнений алгебры предикатов для реализации полученных логических сетей. Результаты диссертационной работы внедрены при создании пакета тендерных соглашений в ООО «ХАРТЕП» (г. Харьков), а также в учебный процесс кафедр АСУ и ИКС НТУ «ХПИ».

Ключевые слова: автоматизированная переработка информации, электронный документооборот, теория категорий, компараторная идентификация, интеллектуальная обработка данных, логические сети.

Syed Mohammad Tauhid Siddiqi. *The models and information technology of intellectual data processing in case of creating business documentation packets.* – Manuscript.

Thesis for a Candidate Degree in Technical Sciences, Specialty 05.13.06 – Information technologies. – National Technical University «Kharkiv Polytechnic Institute», Kharkiv, 2009.

The dissertation is devoted to elaboration of the models and information technology of intelligent data processing for creation of packages of business documentation. An analysis is carried out in research for basic directions of automation of records and creation of packages of business documentation in electronic document-circulation systems, problems of creation of electronic archive of documents and its basic functions, mathematical model of intelligent data processing is developed, allowing passing to intelligent-oriented methods of document processing. The theory of category is chosen as the mathematical tools for simulating of the processes of text data processing within the documents, instrument of logical networks and comparative identification method. A comparative analysis of processing of text information with the help of technologies of Text Mining and comparative identification is carried out. Properties of categories and possibilities of predicate categories for presentation of text information are studied.

The predicate category that is the interpretation of a category within the terminology of the predicate algebra is entered, both cases are considered: categories with objects and without objects. It is shown that the main body of a linear logical operator can be considered as the knowledge or the rules for receiving data and the linear logical operator itself can be considered as the mechanism of fulfillment of inquiry for obtaining of new knowledge. Application of predicate categories for formalization of the knowledge, presented with the help of logical models is displayed. Multiplication of morphisms' properties is examined. Functioning of logical model of representation of knowledge is reproduced in the form of sequence of display and category diagrams. The concept of the modified category is used. The concept of predicate interpretation of a category is generalized.

Intelligent data processing on the basis of application of the theory of categories and logical networks is improved. The algebraic model of identification of documents like as the 'Letter of Enquiry' is offered. Formal signs of the document like as the 'Letter of Enquiry' are selected, areas of their change are defined. The basic types of documents are considered and a logical network of document structure is built according to the type of the commodity.

The model of a logical network is developed for some types of documents which are a part of a package of the documentation under tender agreements. A set of program for solving of the predicate

algebra equations is developed for realization of gained logical networks. Results of dissertational work are adopted in 'Khartep Ltd (Kharkov)' in the time of preparation of tender package and also in educational process of the department 'Automated Control System (ACS)' and 'Intelligent Computer Systems (ICS)' of National Technical University – 'KhPI'.

Key words: automated processing of the information, electronic document-circulation, theory of categories, comparative identification, intelligent data processing, logical networks.

A small, square, light gray image containing a handwritten signature or set of initials in dark ink. The signature is stylized and appears to be written on a white background.