Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

ЄГОРОВ СТАНІСЛАВ ВЯЧЕСЛАВОВИЧ

УДК 004.627:004.912

МОДЕЛІ, МЕТОДИ ТА ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ СЕМАНТИЧНОГО СТИСНЕННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ

05.13.06 – інформаційні технології

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2014

Дисертацією є рукопис.

Робота виконана в Харківському національному університеті радіоелектроніки Міністерства освіти і науки України.

Науковий керівник: кандидат технічних наук, професор
Дудар Зоя Володимирівна,
Харківський національний університет
радіоелектроніки, директор центру
післядипломної освіти

Офіційні опоненти: доктор технічних наук, професор
Каргін Анатолій Олексійович,
Донецький національний університет,
МОН України, м. Вінниця,
завідувач кафедри
комп'ютерних технологій

доктор технічних наук, професор
Асєєв Георгій Георгійович,
Харківська державна академія культури
Міністерства культури України
завідувач кафедри інформаційних
технологій

Захист відбудеться « » _____ 2014 р. о ____ годині на засіданні спеціалізованої вченої ради Д 64.052.08 Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

З дисертацією можна ознайомитися в бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

Автореферат розісланий « » _____2014 р.

Учений секретар
спеціалізованої вченої ради І.П. Плісс

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. В теперішній час стрімкий розвиток Інтернет-технологій, а також баз даних, породжує необхідність зберігання та обробки все більших об'ємів інформації. Зберігання інформації у масштабах Мережі або інформаційно-пошукової системи потребує величезних обчислювальних та часових затрат та у деяких випадках робить інформацію, що надається, застарілою.

Задля скорочення об'ємів інформаційних сховищ, а також з метою прискорення видачі необхідної інформації за запитом та вирішення низки інших актуальних проблем, доцільно застосовувати та вдосконалювати технології стиснення інформації.

Покладений у основу існуючих моделей підхід лише частково описує методологію стиснення тексту. Невирішеним залишається питання контролю та управління семантичним стисненням текстової інформації, а також встановлення залежності семантики від рівня стиснення та об'єму видання. Неможливість інтерактивно впливати на результати стиснення заважає автоматизації усього процесу семантичного стиснення тексту та інтеграції нових методів у сучасні інформаційні системи. Практичні задачі висувають вимоги не тільки власне стиснення, але й оцінки його результатів, а також їх використання в нових, найбільш затребуваних інформаційних технологіях, таких як семантичний та інформаційний пошук, кластерний аналіз, data mining.

Отже, існують проблеми, невирішеність яких призводить до невідповідності стану питання семантичного стиснення текстової інформації сучасним вимогам.

Тому, розробка моделей, методів та інформаційної технології семантичного стиснення текстової інформації, що забезпечують стиснення тексту із заданим рівнем, збереження та оцінку семантичної складової результатів, є актуальною.

Зв'язок роботи із науковими програмами, планами, темами. Дослідження, представлені в дисертаційній роботі, проводилися автором у якості виконавця розділу № 265-3 «Методи та моделі побудови web-сервісів для задач семантичного пошуку текстової інформації в інтелектуальному інформаційному середовищі» фундаментальної НДР № 265 «Методи та моделі самоорганізації інфраструктури інтелектуального інформаційного середовища, що базується на використанні принципів хмарних обчислень», (ДР 0112U000206) згідно із планом науково-дослідних робіт Харківського національного університету радіоелектроніки. В рамках зазначеної НДР здобувачем розроблені моделі, методи та спеціалізована автоматизована система (АС) семантичного стиснення текстової інформації, що дозволяє розв'язати задачу керованого стиснення тексту шляхом задання та корегування рівня стиснення.

Мета та задачі дослідження. Метою дисертаційної роботи є розробка нових, науково-обгрунтованих методів, моделей та інформаційної технології семантичного стиснення текстової інформації.

Для досягнення поставленої мети необхідно розв'язати такі наукові задачі:

– аналіз існуючих моделей та методів семантичного стиснення текстової інформації;

– розробка та дослідження моделі семантичного стиснення текстової інформації з урахуванням рівня стиснення;

– розробка методу семантичного стиснення тексту із заданим рівнем за умови збереження семантичної складової вхідного тексту;

– вдосконалення методу інформаційного пошуку;

– розвиток методу кластеризації текстової інформації;

– розробка інформаційної технології семантичного стиснення текстової інформації для розв'язання задачі контрольованого та керованого стиснення тексту;

– практична апробація та впровадження результатів роботи.

Об'єктом дослідження є процес семантичного стиснення текстової інформації.

Предметом дослідження є моделі, методи та інформаційна технологія семантичного стиснення текстової інформації.

Методи дослідження. Для інформаційного моделювання процесу семантичного стиснення текстових даних застосовується теорія множин та математична логіка; для оцінювання результатів семантичного стиснення - теорія нечітких множин; для розробки методу семантичного стиснення тексту із заданим рівнем - положення дискретної математики та векторної алгебри; для модернізації методів кластеризації, заснованих на щільності, - кластерний аналіз; для розробки програмної реалізації інформаційної технології - теорія проектування архітектури інформаційних систем, шаблони проектування та об'єктно-орієнтований аналіз.

Наукова новизна отриманих результатів. Основні результати, що визначають наукову новизну дисертаційної роботи, полягають у такому:

– вперше розроблено модель семантичного стиснення текстової інформації, яка базується на формалізації функціональних перетворень текстових даних із урахуванням рівня стиснення, що забезпечує реалізацію контрольованого семантичного стиснення текстів будь-яких об'ємів, а також оцінку його результатів;

– вперше розроблено метод семантичного стиснення тексту, який на відміну від існуючих, за рахунок здійснення стиснення із заданим рівнем та встановлення залежності між семантикою та рівнем стиснення, дозволяє здійснювати кероване стиснення текстів, створювати анотації заданого об'єму та контролювати рівень збереження семантики вхідного тексту;

– вдосконалено метод інформаційного пошуку, який на відміну від відомих, за рахунок використання текстових анотацій, отриманих у результаті семантичного стиснення із рівнем, що визначено диференційовано для текстів різних об'ємів, дозволяє підвищити швидкість обробки запиту, покращити умови роботи користувача та надати йому нові можливості в галузі інформаційного пошуку;

– набув подальшого розвитку метод кластеризації текстової інформації, заснований на щільності, який на відміну від існуючих методів, за рахунок використання кластерів максимальної щільності, незалежності від вхідних параметрів та якості вхідних даних, а також застосування паралельної обробки даних, дозволяє здійснювати автоматичну кластеризацію текстової інформації в сховищах даних, а також підвищити точність її обробки.

Практичне значення отриманих результатів. Інформаційна технологія, що базується на розроблених моделях та методах семантичного стиснення текстової інформації реалізована у вигляді спеціалізованої АС та дозволяє здійснювати автоматизоване контрольоване та кероване стиснення тексту за змістом.

Моделі, методи та спеціалізована АС семантичного стиснення текстової інформації були використані для спрощення та прискорення процедури аналізу та обробки текстових документів, скорочення часу вивчення нормативно-технічної документації персоналом; метод інформаційного пошуку - для поліпшення якості та скорочення часу інформаційного пошуку текстових документів у інформаційній системі "Українсько-Словенського Підприємства "Хлібопекарський комплекс "Кулиничівський" (акт від 16.04.2014).

Основні результати дисертаційної роботи використовуються у навчальному процесі Харківського національного університету радіоелектроніки на кафедрі програмної інженерії при проведенні лекційних, лабораторних та практичних занять з дисциплін «Біоніка інтелекту», «Формальні методи програмної інженерії», «Інтелектуальні системи в Інтернет», а також у дипломному та курсовому проектуванні (акт від 30.05.2014), та в держбюджетній НДР №265 (акт від 05.06.2014).

Особистий внесок здобувача. Усі наукові результати дисертаційної роботи, що виносяться на захист, отримані автором самостійно. Праці [4, 8, 11, 12, 13, 14, 15, 16, 17] опубліковані одноосібно. У працях, опублікованих у співавторстві, автору належать такі результати: [1] – вдосконалені методи кластерного аналізу: DBSCAN, OPTICS, K-means и K-medoid; [2] – оптимізація низки алгоритмів кластерного аналізу та їх програмна реалізація; [3] систематизація семантично залежних та семантично незалежних методів стиснення текстової інформації та «Принципу Кількості Коду»; [5] – розробка методу поетапного інформаційного пошуку на основі бази даних анотацій; [6] – розробка моделі семантичного стиснення текстової інформації, що здійснює кероване стиснення тексту з урахуванням його семантичної складової; [7] – розробка методу семантичного стиснення тексту із заданим рівнем; [9] – розробка методу семантичного стиснення тексту із заданим рівнем; [10] – розробка автоматизованої системи семантичного анотування тексту та її програмна реалізація.

Апробація результатів дисертації. Основні положення та результати дисертаційної роботи були представлені, докладались та обговорювались на наступних форумах та семінарах: 14-му Міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI сторіччі» (Харків, 2010); 15-му Міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI сторіччі» (Харків, 2011); 16-му Міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI сторіччі» (Харків, 2012); Всеросійській науковій конференції молодих вчених: «Інновації молодіжної науки» (Санкт-Петербург, 2013); 17-му Міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI сторіччі» (Харків, 2013).

Публікації. За результатами дисертаційних досліджень опубліковано 17 наукових праць, з них: 7 статей у періодичних виданнях України з технічних наук, включених до переліків МОН України, в тому числі 3 статті в журналах, що

входять до міжнародних наукометричних баз; 1 стаття у закордонному журналі; 1 патент України, 1 свідоцтво про реєстрацію авторського права; 7 публікацій у працях конференцій.

Структура та об'єм дисертації. Дисертація складається зі вступу, п'яти розділів, висновків, списку використаних джерел зі 129 найменувань (13 стор.) та 3 додатків (на 6 сторінках). Повний обсяг дисертації складає 189 сторінок, містить 43 рисунка та 13 таблиць (1 рисунок займає окрему площу на 1 сторінці).

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтована актуальність обраної теми дисертації, сформульовано мету та задачі дослідження, визначені об'єкт, предмет та методи досліджень, визначені наукова новизна та практичне значення отриманих результатів, а також особистий внесок автора в роботах, що були виконані у співавторстві, наведені відомості щодо апробації результатів дисертації та кількість публікацій за темою дисертаційної роботи.

У першому розділі на базі вивчення літературних джерел встановлено, що семантичний підхід до стиснення текстової інформації істотно відрізняється від фізичного стиснення тексту та відноситься до незворотних методів стиснення, що не передбачають можливість відновлення вхідного тексту.

Проведене дослідження семантично незалежних методів стиснення, як адаптивних, так і статистичних, дозволило визначити, що стиснення тексту цими методами здійснюється за допомогою кодування ланцюжків символів кодами змінної довжини, а в якості одиниці інформації розглядаються біти або байти, що не дозволяє використовувати їх для збереження сенсу тексту.

Аналіз семантично залежних методів стиснення тексту дозволив встановити, що перетворення синтаксично складних речень в прості може призводити до викривлення сенсу вхідного тексту. Визначено, що з метою збереження семантичної складової тексту слід у якості лексичної одиниці розглядати слова та речення.

Встановлено, що методи автоматичного узагальнення тексту, що дозволяють спростити процес обробки документів великих об'ємів, не можуть самостійно витягувати з вхідного тексту найбільш навантажені за сенсом семантичні одиниці та потребують використання допоміжних інструментів.

Проведене дослідження дозволило встановити, що в існуючих методах стиснення текстової інформації відсутня можливість контролю та керування процесом стиснення, а також оцінки збереження семантичної складової його результатів.

На основі проведеного аналізу проблем, моделей, методів та перспектив семантичного стиснення тексту сформульовано мету та задачі дослідження, що полягають у розробці моделей, методів та інформаційної технології семантичного стиснення текстової інформації.

У другому розділі розроблено модель семантичного стиснення текстової інформації, що вперше надає користувачу можливість задання та коригування рівня стиснення. При цьому, модель гарантує збереження семантичної складової вхідного

тексту. В моделі введено новий компонент, що характеризує рівень стиснення тексту *CR*. Таким чином, модель семантичного стиснення тексту представлена у вигляді кортежу

$$M^{SC} =< X, Y, F, CR >, \quad (1)$$

де $X =< ST_x, W_x >$ – множина вхідних даних, представлена упорядкованим набором речень у вхідному тексті ST_x, а також кортежем слів вхідного тексту W_x; $Y =< ST_y, W_y >$ – множина вихідних даних, представлена упорядкованим набором речень ST_y, що увійшли до анотації, та упорядкованим набором слів анотації W_y; $F = \{F_i | i = \overline{1,T}\}$ – множина функцій перетворення вхідних даних у вихідні; $CR = \{cr_i | i = \overline{1,B}\}$ – множина значень рівня стиснення.

Кортежі слів вхідного тексту W_x та тексту анотації W_y представлені в моделі (1) відповідними множинами $W_x = \{KW_x, SW_x, OW_x\}$ та $W_y = \{KW_y, SW_y, OW_y\}$, що містять множини ключових слів KW_x, KW_y, множини стоп-слів SW_x, SW_y та множини звичайних слів OW_x, OW_y відповідно.

Визначені функціональні перетворення даних. Уведені поняття рівня стиснення тексту, рангу слова, модифікованого інвертованого індексу (МІІ) та семантичної фільтрації тексту.

Задано внутрішнє представлення тексту в моделі (1), що отримано в результаті перетворювання $W_x \xrightarrow{F} W_y$. Такий підхід дозволяє видалити з тексту «шумову» складову у вигляді стоп-слів та здійснити групування слів за їх основами у вигляді модифікованого інвертованого індексу, кожний елемент якого представлено трійкою $e_i = \langle w_i'', r_i, n_i \rangle$, де w_i'' - основа слова для *i*-го елементу індексу, r_i – ранг *i*-го елементу індексу, n_i - номера речень, у яких зустрічається *i*-а основа. Наступні упорядкування елементів індексу та відбір найбільш семантично навантажених речень дозволяє отримати в результаті стиснення текст, об'єм якого залежить від заданого користувачем рівня стиснення.

Розроблена модель надає можливість здійснення контрольованого стиснення текстів будь-яких об'ємів та дозволяє підвищити точність та швидкість їх обробки.

Для оцінки результатів стиснення тексту було розроблено нечітку модель системи оцінки результатів семантичного стиснення, що містить вхідні лінгвістичні змінні «рівень стиснення» та «об’єм видання», та вихідну лінгвістичну змінну «семантика». Кожна з лінгвістичних змінних задана параметрами

$$\langle T, U, G, M \rangle,$$

де T – терм-множина, кожний елемент якої задається нечіткою множиною на універсальній множині U; G – синтаксичні правила, що породжують назви термів; M – семантичні правила, що задають функції приналежності нечітких термів, що породжені синтаксичними правилами G.

Для лінгвістичної змінної «рівень стиснення» визначено універсальну множину $U = [0,100)$ (процентів), терм-множину T = {«низький», «середній», «високий»,

«найвищій»}, синтаксичні правила G та семантичні правила M, що породжують функції приналежності нечітких підмножин універсальної множини U.

Метод побудови функцій приналежності засновано на статистичній обробці думок групи експертів.

Функція приналежності терму «низький» є Z-подібною, а терму «найвищій» - S-подібною

$$\mu_{низький}(u) = \begin{cases} 1, u \le 20 \\ \frac{1}{2} + \frac{1}{2}\cos\left(\frac{u-20}{10}\pi\right), 20 \le u \le 30 \\ 0, u > 30 \end{cases}$$

$$\mu_{найвищій}(u) = \begin{cases} 0, u < 75 \\ \frac{1}{2} + \frac{1}{2}\cos\left(\frac{u-85}{10}\pi\right), 75 \le u \le 85 \\ 1, u > 85 \end{cases}$$

Функції приналежності термів «середній» та «високий» - дзвоноподібні

$$\mu_{середній}(u) = \frac{1}{1 + \left|\frac{u-40}{15}\right|^8}$$

$$\mu_{високий}(u) = \frac{1}{1 + \left|\frac{u-70}{15}\right|^6}$$

Вхідна лінгвістична змінна «об'єм видання» задана на універсальній множині $U = (0,1000]$ (сторінок), терм-множині T = {«малий», «середній», «великий», «найбільший»}, синтаксичними правилами G та семантичними правилами M.

Побудовані функції приналежності термів «малий» - Z-подібна, «середній» та «великий» - дзвоноподібні, «найбільший» - S-подібна зі своїми коефіцієнтами відповідно.

Вихідна лінгвістична змінна «семантика» задана терм-множиною T = {«збережена меншою мірою», «збережена», «збережена повністю»}, кожний елемент якої задано нечіткою множиною на універсальній множині $U = (0,1]$. Для кожного з термів побудовані функції приналежності, що відповідають сігмоїдним та сінглтонній функціям

$$\mu_{збережена\ меншою\ мірою}(u) = \frac{1}{1 + exp^{20(u-0,15)}}$$

$$\mu_{збережена}(u) = \frac{1}{1 + exp^{-30(u-0,5)}}$$

$$\mu_{збережена\ повністю}(u) = \begin{cases} 1, u = 1 \\ 0, u \ne 1 \end{cases}$$

Для всіх лінгвістичних змінних синтаксичні правила, що породжують нові терми, визначено в роботі за допомогою квантифікаторів «ні», «дуже», «більш-менш», а семантичні правила реалізовано за допомогою операцій концентрації та розтягування.

Побудовано нечітку базу знань Мамдані. На основі сформульованих лінгвістичних правил бази знань машина нечіткого логічного виводу визначає значення вихідної змінної у вигляді нечіткої множини $\tilde{Y}$. Проведена дефазифікація, що здійснена в моделі за методом центру тяжіння, перетворює нечітку множину $\tilde{Y}$ у чітке число Y.

Створена нечітка модель системи оцінювання результатів семантичного стиснення тексту встановлює залежність вихідної лінгвістичної змінної «семантика» від двох вхідних лінгвістичних змінних «рівень стиснення» та «об’єм видання» та дозволяє отримати по заданих числових значеннях вхідних змінних чітке значення показника збереження семантики (рис.1)

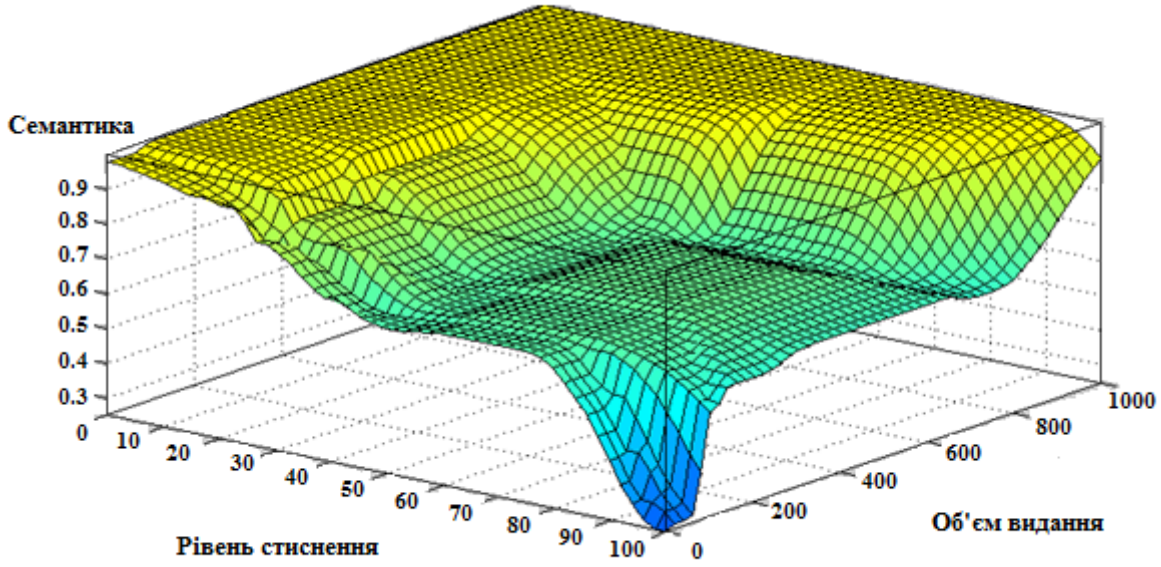

Рисунок 1 - Поверхня виводу нечіткої моделі системи оцінки результатів семантичного стиснення тексту

У третьому розділі розроблено метод семантичного стиснення тексту із заданим рівнем, який використовує ряд понять, що було введено в моделі (1), та визначає ряд нових, таких як: ранги речень та тексту, відносний ранг слова, інвертований індекс анотації (ІІА).

Запропоновано поетапну реалізацію методу.

І етап. Проведення семантичної фільтрації вхідного тексту включає процедури видалення стоп-слів та стемінгу. В результаті першого етапу отримуємо набір основ слів.

ІІ етап. Формування модифікованого інвертованого індексу передбачає створення списку основ слів, що містить інформацію про ранги та відносні ранги основ, а також номери речень, у яких ці основи зустрічаються. Основи слів впорядковано за спаданням їх відносних рангів.

III етап. Створення інвертованого індексу анотації здійснюється шляхом вибору з МІІ ключових слів, починаючи з першого, що мають найвищі показники відносних рангів $R_{\text{відн i}}$. При цьому, на кожному кроці розраховується відносний ранг анотації $R_{\text{відн sum}} = \sum_{i=1}^{N_{sum}} R_{\text{відн i}}$, де N_{sum} – кількість ключових слів у анотації. Включення до розглядання кожного нового ключового слова здійснюється відповідно до перевірки умови з урахуванням коефіцієнту стиснення K

$$R_{відн\ sum} \geq (1 - K)$$

IV етап. Створення анотації передбачає роботу з ІІА, що містить список ключових слів, які мають найвищі показники $R_{\text{відн i}}$, та номери речень, в яких ці слова зустрічаються, задля формування таблиці рангів речень S_j. До анотації будуть включені речення, які мають найвищі показники рангів S_j, тобто ті, що містять найбільшу кількість ключових слів. Кількість речень P_{sum}, які увійдуть до анотації, визначається згідно умови з урахуванням кількості речень вхідного тексту P

$$P_{sum} \geq P\ x\ (1 - K)$$

Метод передбачає можливість багаторазового завдання та коригування рівня стиснення вхідного тексту користувачем (рис.2)

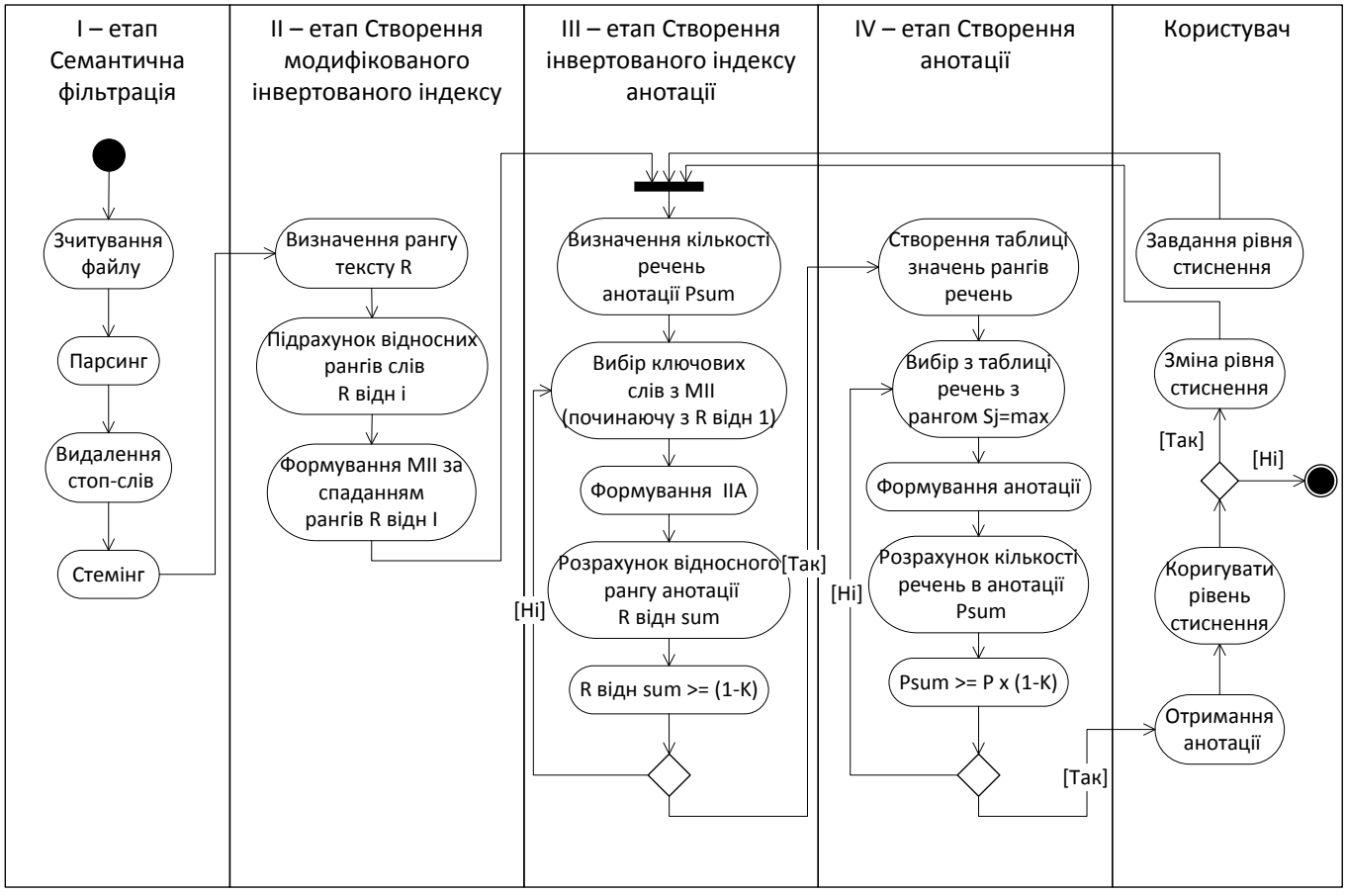

Рисунок 2 – Етапи реалізації методу семантичного стиснення тексту із заданим рівнем

Реалізація методу здійснена таким чином, що коригування рівня стиснення впливає на роботу лише двох останніх етапів методу, тобто створення ІІА та анотації. Такий підхід дозволяє виключити повторне виконання найбільш трудомістких процедур. Формування анотації відбувається із урахуванням не тільки кількісного, але й семантичного аспекту, що забезпечує включення до анотації речень, які найбільш повно відображують сенс вхідного тексту

В методі встановлено залежність між семантикою та кількісними показниками стиснення.

Здійснено оцінку рівня збереження семантики вхідного тексту в анотації, що її було отримано в результаті стиснення (рис.3)

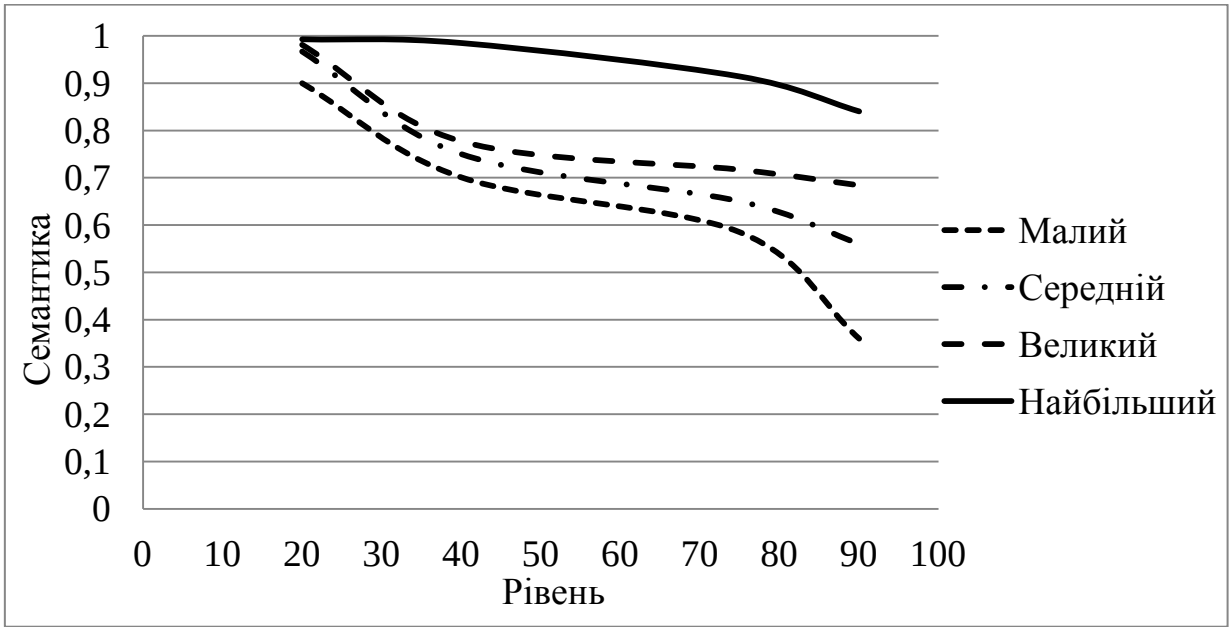

Рисунок 3 – Результати дефазифікації для видань різних об'ємів

Проведена дефазифікація значень вихідної змінної «семантика» засвідчує тенденцію падіння рівня збереження семантики із збільшенням рівня стиснення та збільшення рівня збереження семантики із збільшенням об'єму видання при фіксованих значеннях рівня стиснення.

Метод апробовано на текстах великих об’ємів та повнотекстових документах із вибірок Reuters-21578 та Reuters-RCV1, що підтвердило його роботоспроможність та ефективність застосування для семантичного стиснення текстів різних об’ємів.

Вперше розроблений метод семантичного стиснення текстової інформації із заданим рівнем (рис.2) дозволяє здійснювати кероване стиснення текстових даних та створювати анотацію заданого об’єму за умови збереження семантичної складової вхідного тексту. Метод використано при розробці спеціалізованої АС семантичного стиснення тексту.

Вдосконалено метод інформаційного пошуку на основі текстових анотацій, відмінностями якого є:

– направлення запиту користувача не у БД документів, а у БД анотацій, сформовану одночасно із основною БД у фоновому режимі;

– анотування документів здійснюється згідно з методом семантичного стиснення тексту із заданим рівнем (рис.2);

– розробка та включення етапу «тонкого налаштування» методу задля визначення припустимого рівня стиснення документів;

– використання розробленої нечіткої моделі системи оцінки результатів семантичного стиснення (рис.1) та її дефазифікація задля визначення значень припустимих рівнів стиснення документів різних об’ємів, що забезпечують збереження заданого рівня семантики;

– можливість отримання користувачем SERP, що містить анотації, та отримання повнотекстових документів за запитом.

Основні етапи реалізації методу:

– І етап. Генерація користувачем запиту та направлення його до БД анотацій, сформовану згідно з результатами виконання етапу «тонкого налаштування» методу та дефазифікації нечіткої моделі;

– ІІ етап. Здійснення пошуку найбільш релевантних запиту документів у БД анотацій;

– ІІІ етап. Генерація SERP та видача її користувачу у вигляді списку, який містить анотації;

– IV етап. Видача користувачу повнотекстових версій документів із основної БД за запитом (рис.4).

Вдосконалений метод інформаційного пошуку на основі текстових анотацій дозволяє підвищити швидкість обробки запиту, покращити якість роботи користувача та надати йому нові можливості в області інформаційного пошуку.

Проведено дослідження методів та алгоритмів кластерного аналізу з метою здійснення автоматичної кластеризації текстових документів у сховищах даних великих об’ємів.

Дослідження найбільш ефективних груп методів кластерного аналізу, таких як методи розділення та методи, засновані на щільності, показало, що вони володіють низкою обмежень. Для алгоритмів K-means та bisecting K-means (методи розділення) необхідно попередньо задати кількість кластерів, яку може бути визначено лише емпіричним шляхом, що при аналізі великого обсягу інформації практично неможливо. Для алгоритмів DBSCAN и OPTICS (методи, засновані на щільності) необхідно завдання користувачем мінімальної кількості членів кластеру, а також визначення радіусу пошуку об’єктів у загальній множини. Користувач повинен задавати ці значення вручну, що впливає на ефективність роботи алгоритму в цілому, оскільки визначення вхідних параметрів залежить від його кваліфікації. Окрім цього, для алгоритму OPTICS необхідно використовувати інші алгоритми, що здійснюють безпосередню кластеризацію упорядкованих об’єктів, а також знаходять загальні для всіх документів колекції терміни.

Відмінними особливостями методу, що набув подальшого розвитку, є:

– використання у якості текстових об’єктів анотацій, які сформовані згідно із методом семантичного стиснення текстової інформації із заданим рівнем;

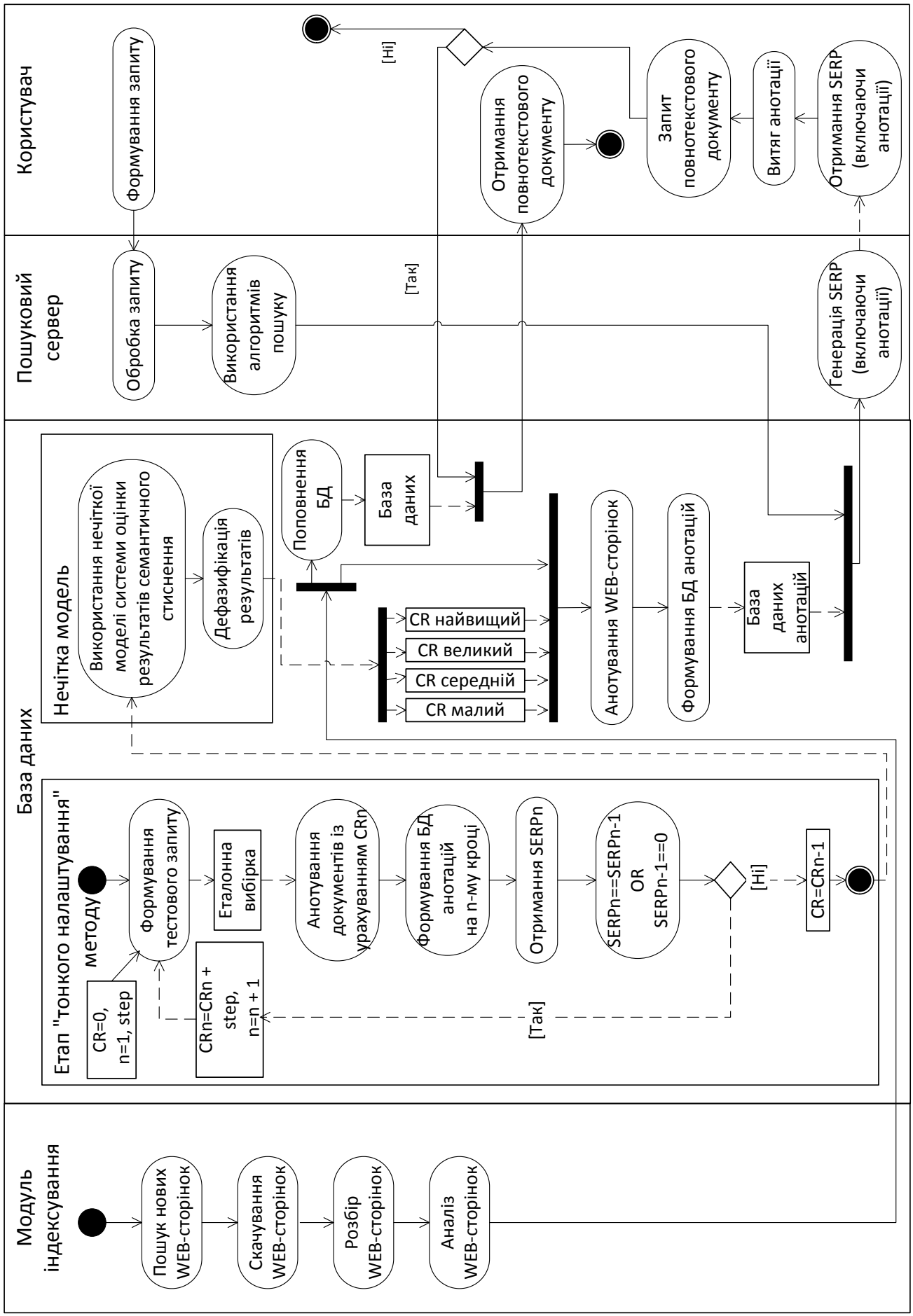

Рисунок 4 – Реалізація вдосконаленого методу інформаційного пошуку

– використання замість інвертованих індексів документів інвертованих індексів анотацій, запропонованих у методі (рис.2), терміни в яких розташовані в порядку спадання їхніх рангів, тобто ваг, що дозволяє аналізувати найбільш значущі для кожного документу терміни;

– формування з анотацій максимально щільних кластерів на основі урахування середньостатистичної кількості слів запиту;

– зняття обмеження на значення мінімальної кількості членів кластеру ($MinPts = 1$);

– використання косинусної міри подібності замість використовуваних у кластерному аналізі метрик задля обчислення радіусу пошуку об'єктів у загальній множині, що для двох документів d_1 та d_2 та відповідним векторним представленням $\vec{\mathrm{V}}(\mathrm{d}_1)$ и $\vec{\mathrm{V}}(\mathrm{d}_2)$ визначається як $\mathrm{sim}(\mathrm{d}_1, \mathrm{d}_2) = \frac{\left(\vec{\mathrm{V}}(\mathrm{d}_1), \vec{\mathrm{V}}(\mathrm{d}_2)\right)}{\|\vec{\mathrm{V}}(\mathrm{d}_1)\|\|\vec{\mathrm{V}}(\mathrm{d}_2)\|}$. Міру може бути визначено автоматично;

– розпаралелювання потоків даних на етапах формування кластерів.

Метод реалізовано у вигляді послідовності етапів.

1 етап. Призначення ядра кластеру з набору об'єктів. У якості об'єктів замість повнотекстових документів розглядаються вектори їх анотацій. Вектори формуються з термінів, що увійшли до ІІА. Кількість таких термінів дорівнює середньостатистичній кількості слів пошукового запиту.

2 етап. Формування максимально щільного кластеру шляхом приєднання об'єктів із максимальним значенням косинусної міри подібності.

3 етап. Формування більш розрідженого кластеру навколо максимально щільного шляхом приєднання об'єктів з меншим значенням косинусної міри подібності.

4 етап. Призначення нового ядра кластеру та повторення етапів 1-3 до кінцевого розподілення усіх об'єктів по кластерах.

Розроблений метод кластеризації текстової інформації дозволяє здійснювати автоматичну кластеризацію текстів у сховищах даних будь-яких об'ємів, пришвидшити сам процес кластеризації та підвищити якість та точність їх обробки.

У четвертому розділі розроблено інформаційну технологію семантичного стиснення текстової інформації, що базується на запропонованих моделях та методах (рис.5).

І стадія реалізує задачі формування функціональних вимог до спеціалізованої АС семантичного стиснення тексту на основі аналізу процесу стиснення з використанням існуючих моделей та методів. Результатом роботи етапу є вимоги користувача до АС, а саме: формати файлів, що підлягають обробці, можливості задання та коригування рівня стиснення, отримання анотації заданого обсягу, а також збереження анотації у вигляді текстового файлу.

ІІ стадія. Розробка концепції АС передбачає перш за все визначення предметної галузі, елементами якої є множини вхідних та вихідних даних, а також множина значень рівня стиснення. Для реалізації виконання функціональних вимог І етапу необхідно розробити моделі та методи семантичного стиснення тексту.

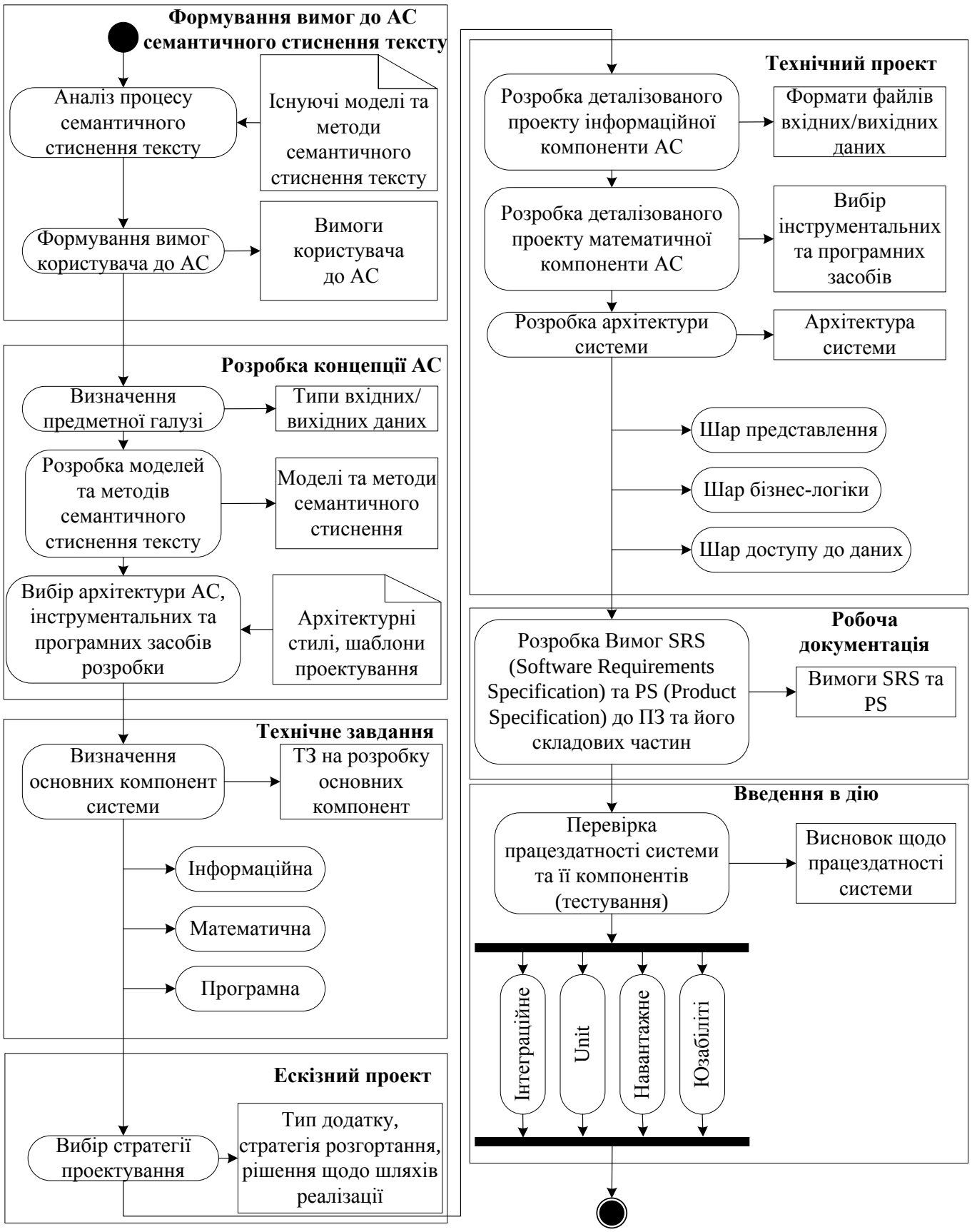

Рисунок 5 – Схема інформаційної технології семантичного стиснення текстової інформації

На основі аналізу предметної галузі та розроблених вимог здійснено вибір архітектури АС, її інструментальних та програмних засобів розробки з урахуванням архітектурних стилів та шаблонів проектування.

На III стадії визначено основні компоненти АС у вигляді інформаційної, математичної та програмної. Запропоновано технічне завдання на розробку основних компонент АС.

В рамках IV стадії здійснено вибір стратегії проектування та отримано ескізний проект.

На V стадії проведена розробка деталізованого проекту інформаційної та математичної компонент АС, а також архітектури системи. Результатом виконання етапу є технічний проект.

В рамках створення робочої документації на VI стадії розроблені вимоги SRS (Software Requirement Specification) та PS (Product Specification) до ПЗ АС.

На VII стадії здійснено перевірку працездатності системи та компонентів. Результатом роботи етапу є введення АС до експлуатації.

Розроблена інформаційна технологія дозволяє здійснювати контрольоване та кероване стиснення тексту, а також автоматизувати процес отримання анотації заданого об’єму.

У п’ятому розділі викладено результати впровадження інформаційної технології у вигляді спеціалізованої АС семантичного стиснення тексту та проведено оцінку її ефективності.

Інтеграція спеціалізованої АС в інтелектуальне інформаційне середовище дозволила винести найбільш трудозатратні обчислення до хмари, спричинила реалізацію шару бізнес-логіки (Business Logic Layer, BL) у вигляді Cloud-сервісу та здійснена з метою захисту ядра технології у вигляді розроблених моделей та методів. Використання технології хмарних обчислень дозволило знизити вимоги, що пред'являються до клієнтських комп'ютерів, обмеживши їх наявністю підключення до Інтернет, та істотно скоротило робоче навантаження та дозволило скоротити кількість робочих серверів з 3 до 1. При цьому, шар інтерфейсу користувача (UI Layer) та шар доступу до даних (Data Access Layer) не піддалися істотним змінам за рахунок використання при проектуванні архітектури АС максимального рівня абстракції між шарами. Інтеграція спеціалізованої АС дозволила надати нову можливість в обробці текстової інформації у вигляді керованого стиснення тексту, істотно скоротити час аналізу та обробки тексту (для текстів об’ємом до 30 сторінок час скорочується з 72 хв. до 4 хв. за рахунок використання спеціалізованої АС замість анотування вручну. При збільшенні об’єму текстів час скорочується багатократно та нелінійно), знизити витрати на адміністрування на 40% та підвищити рівень безпеки системи, а також скоротити час виконання запитів (в середньому з 15с до 4с) та підвищити якість інформаційного пошуку.

Проведене тестування інтегрованих програмних засобів підтвердило ефективність застосування розробленої інформаційної технології задля вирішення задач семантичного стиснення та пошуку текстової інформації.

На підприємстві "УСП "Хлібопекарський комплекс "Кулиничівський" вдосконалено корпоративну інформаційну систему за рахунок впровадження

розроблених методів та АС семантичного стиснення текстової інформації. В рамках впровадження методу інформаційного пошуку було створено альтернативну базу даних анотацій із застосуванням алгоритму «тонкого налаштування», що дозволило скоротити час інформаційного пошуку текстових документів у інформаційній системі підприємства з 10с до 3с. Розроблено інтерфейс користувача, що дозволяє максимально спростити дії користувача по роботі з системою.

За рахунок впровадження спеціалізованої АС семантичного стиснення тексту на підприємстві користувачам було надано якісно нові можливості по роботі з документацією, спрощено процедуру обробки текстових документів, прискорено їх аналіз та скорочено час вивчення нормативно-технічної документації персоналом підприємства.

У додатках наведено акти про впровадження результатів дисертаційної роботи у виробництво, держбюджетну НДР та навчальний процес.

ВИСНОВКИ

В дисертаційній роботі розв’язано важливу науково-технічну задачу, а саме розробку моделей, методів та інформаційної технології семантичного стиснення текстової інформації, що дозволяє здійснювати кероване стиснення вихідного тексту за допомогою задання рівню стиснення та контролювати збереження семантичної складової його результатів.

В рамках проведеного дослідження отримані такі наукові результати:

1. Проведено аналіз моделей та методів семантичного стиснення текстової інформації з урахуванням сучасних вимог обробки, зберігання та передачі інформації, який підтвердив актуальність вирішення задачі стиснення тексту із заданим рівнем для реалізації можливості контролю та керування процесом семантичного стиснення тексту.
2. Вперше запропоновано модель семантичного стиснення текстової інформації, що базується на формалізації функціональних перетворень текстових даних із урахуванням можливості задання та корегування рівня стиснення. Дано визначення рівня стиснення тексту, рангу слова, модифікованого інвертованого індексу, семантичної фільтрації тексту, а також внутрішнього представлення даних. Модель семантичного стиснення тексту може бути використана задля здійснення контрольованого стиснення текстів будь-яких об’ємів.
3. Розроблено нечітку модель системи оцінки результатів семантичного стиснення тексту, у якій встановлено залежність рівня збереження семантики від рівня стиснення та об’єму вхідного тексту. Модель дозволяє отримати чітке значення рівня збереження семантики при завданні значень рівня стиснення та об’єму тексту.
4. Вперше розроблено метод семантичного стиснення тексту, що на відміну від існуючих, здійснює стиснення із заданим рівнем та встановлює залежність між семантикою та рівнем стиснення. Введені поняття відносного рангу слова, рангу речення, інвертованого індексу анотації. Метод дозволяє здійснювати

кероване стиснення текстів, створювати анотації заданого об’єму та контролювати рівень збереження семантики вхідного тексту.

5. Вдосконалено метод інформаційного пошуку, який на відміну від відомих, за рахунок використання текстових анотацій, отриманих у результаті семантичного стиснення із припустимим рівнем, дозволяє підвищити якість роботи користувача, надати йому нові можливості в області пошуку інформації та істотно підвищити швидкість обробки запиту. Реалізовано алгоритм етапу «тонкого налаштування» методу та розроблено диференційований підхід до визначення припустимих рівнів стиснення документів різних об’ємів, що дозволяє зберегти заданий рівень семантики.

6. Набув подальшого розвитку метод кластеризації текстової інформації, заснований на щільності. Метод відрізняється використанням кластерів максимальної щільності, незалежністю від вхідних параметрів та якості вхідних даних, а також застосуванням паралельної обробки інформації. Метод дозволяє здійснювати автоматичну кластеризацію текстів у сховищах даних, а також підвищити точність її обробки.

7. Розроблено інформаційну технологію семантичного стиснення текстової інформації на основі запропонованих у роботі моделей та методів. Технологія дозволяє автоматизувати процес контрольованого та керованого семантичного стиснення тексту, а також отримання анотацій заданого об’єму. На базі інформаційної технології розроблено спеціалізовану АС семантичного стиснення текстової інформації.

8. Підтверджено достовірність отриманих у дисертаційній роботі результатів проведеною практичною апробацією, а також результатами впровадження на реальних об’єктах.

Розроблені інструментальні та програмні засоби семантичного стиснення текстової інформації можуть бути використані при розв’язанні задач керованого семантичного стиснення текстів, інформаційного пошуку, організації альтернативних сховищ текстових даних, автоматичної кластеризації інформації в таких сховищах, а також її верифікації.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Егоров С. В. Разработка программного обеспечения для решения задач распознавания образов / И. Н. Егорова, С. В. Егоров // Восточно-Европейский журнал передовых технологий. – Х., 2010. – № 1/5 (43). – С. 67–68.

2. Егоров С. В. Оптимизация алгоритмов кластерного анализа в задачах распознавания образов / С. В. Егоров, З. В. Дударь // Вестн. Херсон. нац. техн. ун-та. – Херсон : Изд-во ХНТУ, 2011. – № 2 (41). – С. 217–219.

3. Егоров С. В. Исследование и оптимизация методов сжатия текстовой информации / С. В. Егоров, З. В. Дударь // Вестн. Херсон. нац. техн. ун-та. – Херсон: Изд-во ХНТУ, 2012. – № 1 (44). – С. 65–68.

4. Егоров С. В. Информационный подход в семантическом сжатии текста / С. В. Егоров // Вестн. мол. ученых С.-Петерб. гос. ун-та технологий и дизайна. – СПб., 2013. – Вып. 1 : Естественные и технические науки. – С. 22–27.

5. Егоров С. В. Семантическое аннотирование в информационно-поисковых системах / С. В. Егоров, З. В. Дударь // Бионика интеллекта. – Х., 2013. – № 1 (80). – С. 104–107.

6. Егоров С. В. Математическая модель семантического сжатия текстовой информации / И. Н. Егорова, С. В. Егоров // Системи обробки інформації. – Х., 2013. – Вип. 9 (116). – С. 18–21.

7. Егоров С. В. Метод семантического сжатия текста / И. Н. Егорова, С. В. Егоров // Вісн. Нац. техн. ун-ту «ХПІ». Сер. Математичне моделювання в техніці та технологіях : зб. наук. пр. – Х. : НТУ «ХПІ», 2013. – № 54 (1027). – С. 118–123.

8. Егоров С. В. Информационная технология семантического сжатия текста / С. В. Егоров // Зб. наук. пр. Харк. ун-ту Повітр. Сил. – Х., 2013. – Вип. 4 (37). – С. 113–116.

9. Пат. на корисну модель 82942 Україна, МПК 51 G 06 F 17/21. Спосіб семантичної компресії тексту із заданим рівнем стислості / С. В. Єгоров, І. М. Єгорова. – № u201300978 ; заявл. 28.01.13 ; опубл. 27.08.13, Бюл. № 16. – 8 с.

10. Свідоцтво про реєстрацію авторського права на твір № 51167 Україна. Комп'ютерна програма «Інформаційна система семантичного анотування тексту» / С. В. Єгоров, І. М. Єгорова. – 06.09.13.

11. Егоров С. В. Программная реализация алгоритмов кластеризации в задачах распознавания образов / С. В. Егоров // Радиоэлектроника и молодежь в XXI веке : материалы 14-го междунар. молодеж. форума, 18–20 марта 2010 г. – Х. : ХНУРЭ, 2010. – С. 477.

12. Егоров С. В. Применение классификаторов для распознавания образов / С. В. Егоров // Радиоэлектроника и молодежь в XXI веке : материалы 14-го междунар. молодеж. форума, 18–20 марта 2010 г. – Х. : ХНУРЭ, 2010. – С. 478.

13. Егоров С. В. Оптимизация алгоритмов кластерного анализа / С. В. Егоров // Радиоэлектроника и молодежь в XXI веке : материалы 15-го Юбилейного Междунар. молодеж. форума, 18–20 апр. 2011 г. – Х. : ХНУРЭ, 2011. – Т. 9 : Междунар. конф. «Информационные интеллектуальные системы». – С. 217.

14. Егоров С. В. Исследование методов сжатия текстовой информации / С. В. Егоров // Радиоэлектроника и молодежь в XXI веке : материалы 16-го Междунар. молодеж. форума, 17–19 апр. 2012 г. – Х. : ХНУРЭ, 2012. – Т. 6 : Междунар. конф. «Информационные интеллектуальные системы». – С. 212–213.

15. Егоров С. В. Совершенствование методов информационного поиска / С. В. Егоров // Радиоэлектроника и молодежь в XXI веке : материалы 17-го междунар. молодеж. форума, 22–24 апр. 2013 г. – Х. : ХНУРЭ, 2013. – Т. 6 : Междунар. конф. «Информационные интеллектуальные системы». – С. 265–266.

16. Егоров С. В. Автоматическое обобщение текстов на естественном языке / С. В. Егоров // Радиоэлектроника и молодежь в XXI веке : материалы 17-го

междунар. молодеж. форума, 22–24 апр. 2013 г. – Х. : ХНУРЭ, 2013. – Т. 6 : Междунар. конф. «Информационные интеллектуальные системы». – С. 267–268.

17. Егоров С. В. Семантическое аннотирование в информационном поиске / С. В. Егоров // Инновации молодежной науки : тез. докл. Всерос. науч. конф. мол. ученых / С.-Петерб. гос. ун-т технологий и дизайна. – СПб., 2013. – С. 113.

АНОТАЦІЯ

Єгоров С.В., Моделі, методи та інформаційна технологія семантичного стиснення текстової інформації. – Рукопис.

Дисертація на здобуття вченого ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Харківський національний університет радіоелектроніки Міністерства освіти і науки України, Харків, 2014.

Дисертаційну роботу присвячено розв'язанню актуальної науково-технічної задачі розробки моделей, методів та інформаційної технології семантичного стиснення текстової інформації.

Розроблено модель семантичного стиснення текстової інформації, що забезпечує реалізацію контрольованого семантичного стиснення текстів будь-яких об'ємів, а також оцінку його результатів. Розроблено метод семантичного стиснення тексту із заданим рівнем, що дозволяє здійснювати кероване стиснення текстів, створювати анотації заданого об'єму та контролювати рівень збереження семантики вхідного тексту. Вдосконалено метод інформаційного пошуку, що дозволяє підвищити швидкість обробки запиту, якісно покращити умови роботи користувача та надати йому нові можливості в галузі інформаційного пошуку. Набув подальшого розвитку метод кластеризації текстової інформації, заснований на щільності, що дозволяє здійснювати автоматичну кластеризацію текстової інформації в сховищах даних, а також підвищити точність її обробки.

На основі запропонованих моделей та методів створено інформаційну технологію семантичного стиснення текстової інформації та здійснено її практичну реалізацію у вигляді спеціалізованої АС.

Ключові слова: модель, метод, інформаційна технологія, семантика, рівень стиснення, припустимий рівень стиснення, рівень збереження семантики.

АННОТАЦИЯ

Егоров С.В., Модели, методы и информационная технология семантического сжатия текстовой информации. – Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 – информационные технологии. – Харьковский национальный университет радиоэлектроники Министерства образования и науки Украины, Харьков, 2014.

Диссертационная работа посвящена решению актуальной научно-технической задачи разработки моделей, методов и информационной технологии семантического сжатия текстовой информации с целью применения их в системах обработки

текстовой информации, хранилищах данных, информационно-поисковых системах.

На основе проведенного анализа существующих методов сжатия текстовой информации установлено, что они не позволяют контролировать и управлять процессом сжатия текста, а также оценивать семантическую составляющую полученных результатов.

Разработана и формально описана модель семантического сжатия текстовой информации в виде набора множеств исходных и выходных данных, множества функций преобразования исходных данных в выходные, а также множества значений уровня сжатия. Дано формализованное описание функциональных преобразований текстовых данных с учетом уровня сжатия. Задано внутреннее представление текста в модели. Введены понятия уровня сжатия текста, ранга слова, модифицированного инвертированного индекса и семантической фильтрации текста. Модель реализует возможность контролируемого семантического сжатия текстов любых объемов.

Предложена нечеткая модель системы оценки результатов семантического сжатия текста. Определены и описаны входные лингвистические переменные «уровень сжатия» и «объем издания», а также выходная лингвистическая переменная «семантика». Построена нечеткая база знаний Мамдани. Проведена дефаззификация по методу центра тяжести. Модель устанавливает зависимость уровня сохранения семантики от уровня сжатия и объема исходного текста и позволяет контролировать сохранение семантической составляющей текстов любых объемов при различных значениях уровня сжатия.

Разработан метод семантического сжатия текста с заданным уровнем. Введены понятия рангов предложения и текста, относительного ранга слова, инвертированного индекса аннотации. Формализовано поэтапное преобразование текстовых данных в аннотацию с учетом семантического аспекта, обеспечивающего включение в аннотацию предложений, наиболее полно отражающих смысл исходного текста. В методе установлена зависимость между семантикой и уровнем сжатия. Осуществлена оценка уровня сохранения семантики исходного текста в аннотации, полученной в результате сжатия. Проведена апробация метода на текстах больших объемов, что подтвердило его работоспособность и эффективность для семантического сжатия изданий любых размеров. Разработанный метод семантического сжатия текстовой информации с заданным уровнем позволяет осуществлять управляемое сжатие текстов, создавать аннотации заданного объема и контролировать уровень сохранения семантики исходного текста.

Усовершенствован метод информационного поиска. Предусмотрена функция аннотирования документов и создание базы данных аннотаций. Определен допустимый уровень сжатия текстовых данных. Реализован алгоритм этапа «тонкой настройки». Разработан дифференцированный подход к определению допустимых уровней сжатия документов разных объемов, который обеспечивает сохранение необходимого уровня семантики. Метод позволяет качественно улучшить условия работы пользователя, предоставить ему новые возможности в области поиска информации и существенно повысить скорость обработки запроса.

Получил дальнейшее развитие метод кластеризации текстовой информации, основанный на плотности. Проведен анализ существующих методов кластерного анализа, определены их ограничения в виде необходимости задания пользователем минимального количества членов кластера и радиуса поиска объектов. Установлены отличительные особенности разработанного метода, а именно: поиск документов в базе данных аннотаций, использование максимально плотных кластеров, независимость от входных параметров и качества исходных данных, параллельная обработка данных. Метод осуществляет автоматическую кластеризацию текстов в хранилищах данных и повышает точность их обработки.

Разработана информационная технология семантического сжатия текстовой информации на основе созданных в работе моделей и методов. Технология направлена на автоматизацию процесса получения аннотации заданного объема и позволяет реализовать управляемое семантическое сжатие текста. Разработана специализированная АС, осуществляющая практическую реализацию предложенной технологии.

Полученные теоретические результаты использованы при разработке и внедрении в систему семантического поиска текстового контента в интеллектуальной информационной среде, корпоративную систему предприятия и учебный процесс, что подтверждается актами о внедрении.

Ключевые слова: модель, метод, информационная технология, семантика, уровень сжатия, допустимый уровень сжатия, уровень сохранения семантики.

ANNOTATION

Iegorov S., Models, methods and information technology for semantic compression of textual information. – Manuscript.

Thesis for obtaining scientific degree of Candidate of Technical Sciences on the speciality 05.13.06 – information technologies. – Kharkiv national university of radio electronics of Ministry of education and science of Ukraine, Kharkiv, 2014.

Thesis is devoted to development of models, methods and information technology for semantic compression of textual information.

Developed model for semantic compression of textual information provides implementation of controllable semantic compression of texts of any volume and results evaluation. Developed method for semantic text compression with given ratio allows to carry on managed compression of texts, create annotations of given volume and control semantic retaining level of initial text. Improved method for information retrieval, which allows to increase query processing speed, effectively improve user experience and grant new possibilities in information retrieval to user. Further developed density-based clustering method for textual information allows to accomplish automatic clusterization of textual information in data warehouses and increase processing precision.

Information technology was created on top of suggested models and methods and implemented as specialized automated system.

Keywords: model, method, information technology, semantic, compression ratio, admissible compression ratio, semantic retain level.

Підп. до друку 11.11.14. Формат 60х84 1/16. Спосіб друку – ризографія.
Умов. друк. арк. 1,2. Облік. вид. арк. 1,0. Тираж 100 прим.
Ціна договірна. Зам. № 2-900.

ХНУРЕ, Україна, 61166, Харків, просп. Леніна ,14

Віддруковано в навчально-науковому
видавничо-поліграфічному центрі ХНУРЕ
61166, Харків, просп. Леніна, 14
Свідоцтво про внесення суб'єкта до Державного реєстру
видавців, виготовників і розповсюджувачів видавничої продукції серія ДК № 1409 від 26.06.03 р.