

**МІНІСТЕРСТВО ОСВІТИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Пляшко Наталія Володимирівна

УДК 004.912

**МЕТОДИ АДАПТАЦІЇ НАВЧАЛЬНИХ КУРСІВ У СИСТЕМАХ БЕЗПЕРЕВНОГО
ДИСТАНЦІЙНОГО НАВЧАННЯ**

Спеціальність 8.050103 –
«Інженерія програмного забезпечення»

Автореферат

дисертації на здобуття освітньо-кваліфікаційного рівня
магістр

Київ – 20--

Робота виконана на кафедрі “Автоматизації проектування енергетичних процесів та систем” НТУУ «КПІ» Міністерства освіти України.

Науковий керівник роботи магістра: кандидат технічних наук, доцент
Гагарін Олександр Олександрович
Національний технічний університет України
„Київський політехнічний інститут”

Рецензенти:

Захист відбудеться __ червня 20-- р. о _____ на засіданні ГЕК кафедри АПЕПС
НТУУ «КПІ» аудиторія 4хх__-5

З дисертацією можна ознайомитись у методичному кабінеті кафедри АПЕПС
НТУУ «КПІ» аудиторія 4хх-5.

Автореферат підготовлено та надано для розгляду “__” травня 20-- р.
Робота рекомендована до захисту “__” червня 20-- р.

Завідувач кафедри АПЕПС НТУУ «КПІ»
.доктор технічних наук, професор

Лук’яненко С.О.

Відповідальний за випуск магістрів
кафедри АПЕПС НТУУ «КПІ»
кандидат технічних наук, доцент

Гагарін О.О.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Протягом останніх десятиліть спостерігається швидкий ріст об'ємів текстової інформації, яка зберігається у вигляді електронних документів. Для ефективної роботи з ними вимагаються сучасні інструменти, важливу роль серед яких відіграють різноманітні засоби інформаційного пошуку. Склалася ситуація, за якої фахівці більше часу витрачають на пошук інформації, ніж на її використання за призначенням. Основна мета інформаційного пошуку – це допомогти користувачу знайти інформацію, яка його цікавить. Система відбирає із всієї наявної множини інформації підмножину, задовольняючи користувача. Весь розвиток сучасних методів інформаційного пошуку можна представити як постійне вдосконалення і ускладнення моделі документа і методів його використання. Чим більший об'єм інформації, що видається, тим складніше в ній розібратися. Таким чином постає задача розробки більш ефективних алгоритмів та методів пошуку. Актуальною темою дисертації являється дослідження впливу вибору моделі і її різних характеристик на якість інформаційного пошуку.

Зв'язок роботи з науковими програмами, планами, темами.

Дисертаційна робота магістра виконувалась у НТУУ "КПІ" у відповідності з планом наукових досліджень кафедри АПЕПС.

Метою дослідження є удосконалення моделей і методів представлення текстового документа в системах інформаційного пошуку, які враховують взаємне розташування слів, вивчити їх вплив на якість інформаційного пошуку і способи їх ефективної реалізації з використанням індексних структур та попереднього автоматичного реферування документів.

Для реалізації поставленої мети були сформульовані наступні **завдання дослідження**, що визначили логіку дослідження та його структуру:

- проаналізувати функціональну структуру сучасних інформаційних систем пошуку;
- проаналізувати існуючі моделі і методи інформаційного пошуку;

- дослідити вплив вибору моделі і її різних характеристик на якість інформаційного пошуку;
- удосконалити модель і метод представлення текстового документа та обґрунтувати доцільність використання цього методу;
- розробити програмне забезпечення реалізації моделі та методів представлення текстового документа в системах інформаційного пошуку.

Об’єктом дослідження є програмне забезпечення систем інформаційного пошуку.

Предмет дослідження – програмне забезпечення для реалізації моделей та методів представлення текстових документів в системах інформаційного пошуку.

Методи дослідження: розв’язання поставлених задач виконувалось на базі дослідження існуючих методів та моделей для представлення текстового документа в системах інформаційного пошуку, зокрема:

- моделі “множини слів” та використання пар слів для представлення текстового документа;
- реферування документів для створення нової скороченої версії документа, щоб скоротити час пошуку;
- зважування реферованого документа за варіантом класичного алгоритму пошуку.

Наукова новизна одержаних результатів. Найбільш суттєвими науковими результатами магістерської дисертації є:

- удосконалено модель та метод представлення текстового документа за рахунок попереднього реферування документа та генерування власної індексної структури, зручної у використанні системи під час пошуку, що зменшує витрати у перерахунку на один документ і дозволяє значно скоротити час пошуку документів;
- набуло подальшого розвитку застосування моделей та методів представлення текстового документа в системах інформаційного пошуку для текстових колекцій та оцінка якості пошуку.

Практичне значення одержаних результатів визначається тим, що запропонований метод пошуку може бути з успіхом застосований у сучасних інформаційних та інформаційно-аналітичних системах для оброблення великих масивів документів. Попереднє автоматичне реферування документів зменшує витрати у перерахунку на один документ і дозволяє значно скоротити час пошуку документів.

Апробація результатів дисертації

Основні положення роботи доповідались і обговорювались на:

1. VII Науково-практичному семінарі з міжнародною участю ім. проф. І.В.Недіна “Економічна безпека держави і науково-технологічні аспекти її забезпечення” (м. Київ, 21-22 жовтня 2015 р).
2. XIV Міжнародній науково-практичній конференції аспірантів, магістрантів і студентів “Сучасні проблеми наукового забезпечення енергетики” (м. Київ, 18-21 квітня 2016 р).
3. III науково-практичній дистанційній конференції молодих вчених і фахівців з розробки програмного забезпечення “Сучасні аспекти розробки програмного забезпечення” (м. Київ, 15 квітня 2016 р.).

Публікації. Наукові положення дисертації опубліковані у 2 роботах.

Ключові слова. ІНФОРМАЦІЙНИЙ ПОШУК, ІНДЕКСНА СТРУКТУРА, ДОКУМЕНТ, ВАГА СЛОВА, ЧАСТОТА СЛОВА, РЕФЕРУВАННЯ, МОДЕЛЬ, $TF*IDF$.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** описано проблему засобів інформаційного пошуку, розписано основні питання, що розглядалися в дослідженні.

У **першому розділі** коротко охарактеризовано контекст дослідження. Зокрема, обґрунтовано актуальність теми дисертаційного дослідження, визначено мету, об’єкт, предмет і основні завдання дисертації. Перераховані і коротко описані основні задачі інформаційного пошуку. Сформульовано визначення моделі документа в системі інформаційного пошуку.

У **другому розділі** розглянуто пошукові системи в основному з точки зору використовуваних в них інформаційних технологій, сформульовано основні поняття та визначення. Головним завданням будь-якої інформаційно-пошукової системи є пошук інформації релевантної інформаційним потребам користувача. Тому розглянуто основні інформаційні потреби користувача, описано чотири головні етапи представлення інформаційної потреби, представлено еволюцію представлень інформаційної потреби.

Описуються підходи, критерії та методи оцінки якості інформаційного пошуку. Під ефективністю пошуку розуміється час виконання запиту. Система при цьому виконує ряд дій: відбір документів, оцінка їх ваги, формування результатів для подання користувачеві. Ефективність будь-якої інформаційної системи визначається її здатністю служити тим цілям, для яких вона була розроблена. В основі всіх методів оцінки лежить поняття релевантності, тобто відповідності документа запиту. Факт відповідності має суб'єктивний характер, так як може бути встановлений тільки людиною. Для отримання більш об'єктивних характеристик оцінку відповідності дають кілька людей і результат усереднюється.

Важливо відзначити, що в результаті аналізу статистики запитів великого числа користувачів по відомим пошуковим системам з'ясувалося, що середня довжина запиту не перевищує двох слів, і користувачі, як правило, використовують найпростішу форму запиту. Тому не можна розраховувати на те, що користувач пошукової системи буде формулювати досить складні ефективні запити. Сучасна пошукова система повинна самотійно виявляти інформаційні потреби конкретного користувача і враховувати їх при пошуку.

У даному розділі наведена загальна схема процесу пошуку (рисунок 1).

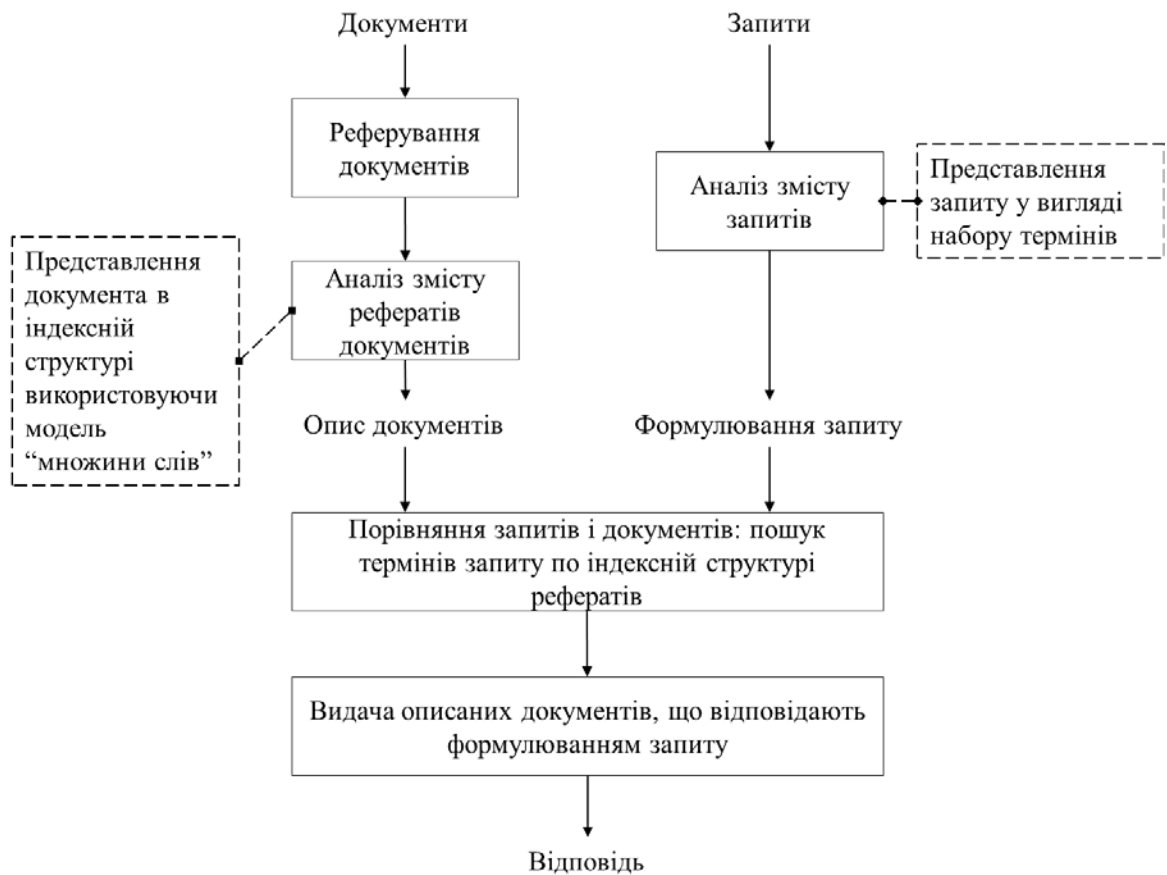


Рисунок 1 – Загальна схема процесу пошуку

Зміст документів і запитів в інформаційно-пошукових системах зазвичай описується деякими наборами термінів, що складаються з окремих слів або словосполучень. Часто для характеристики термінів використовуються їх вага, які відображають передбачувану важливість кожного з термінів. Рішення про видачу того чи іншого документа приймається в результаті порівняння наборів термінів, які використовувалися відповідно до документів і запитів. До уваги користувача пропонуються ті документи, набори термінів яких співпадають з наборами термінів запитів.

У **третьому розділі** приводиться огляд відомих моделей документів в системах інформаційного пошуку. Існує безліч моделей представлення документа при інформаційному пошуку. В даному розділі робиться висновок, що історично перша і найпростіша модель множини слів (bag of words) поступилася своїм місцем більш складним, але дозволяє досягти підвищення якості інформаційного пошуку.

Найбільш поширеними і відомими, які дають велике збільшення якості при відносно малих накладних витратах і простоті реалізації є:

- документ, як множина ваг термінів;
- документ, як множина фрагментів;
- документ, як вузол гіпертекстового графа.

Саме ці моделі, а точніше, їх комбінації, знаходять застосування в комерційних і дослідницьких системах.

Передбачено перспективи моделей і методів інформаційного пошуку. Якість інформаційного пошуку сучасних систем неухильно зростає. Це обумовлено зростаючими обчислювальними потужностями та обсягами пам'яті сучасних комп'ютерів, що дозволяє реалізовувати все більш складні і ресурсомісткі алгоритми. Багато описаних моделей, особливо пов'язані зі складним аналізом тексту, такі як синтаксичний аналіз, поки просто неможливо реалізувати в повному обсязі, тому важко оцінити їх вплив на якість пошуку. Інші поки що недостатньо вивчені з точки зору їх впливу на якість пошуку і вимагають додаткових досліджень. Не можна сказати, що існуючі моделі подання документа дозволяють використовувати для інформаційного пошуку всю інформацію, яку можна отримати з документа. Очевидно, що можна створити нові та/або використовувати комбінації вже існуючих моделей.

У **четвертому розділі** наведено підходи та методи представлення текстового документа в системах інформаційного пошуку: індексування, реферування документів. Інформаційно-пошукова система для швидкого пошуку та економії дискового простору повинна мати компактне представлення моделі документа.

Перше завдання, яке повинна вирішити інформаційно-пошукова система – це створення опису документа або будь-якого іншого інформаційного ресурсу. Така процедура називається індексуванням. Основна мета процесу індексування – поставити у відповідність кожному документу деяку множину ключових слів, які відображають зміст документа. Незважаючи на високу швидкодію сучасних комп'ютерів, неможливо забезпечити обробку запиту за прийнятний час для

великих колекцій документів без використання спеціальних індексних структур. Спостережуване постійне зростання розмірів колекцій призводить до того, що використовувані методи повинні добре масштабуватися, забезпечуючи зростання часу на виконання запиту із меншою швидкістю, ніж лінійний від розміру колекції.

В даний час запропоновано і досліджено велику кількість різних варіантів індексних структур. В роботі розглянуто ті, які дозволяють реалізувати розглянуті моделі документа, побудовані на основі вхідних у нього термінів.

В даний час для оцінки ваг термінів використовується головним чином наступні три моделі:

- частотна модель;
- імовірнісна модель;
- латентно-семантичний аналіз.

В п'ятому розділі обґрунтовано вибір та описано особливості реалізація моделі та методу представлення документа. Існує ряд підходів до використання інформації про взаємне положення слів. Дані дослідження показують, що ця інформація покращує якість пошуку, але не існує загальноприйнятої думки про те, який з методів її використання найбільш ефективний. У даній роботі досліджувалось два підходи, які здалися найбільш перспективними:

- виділення контактних “пар слів”;
- реферування.

Кожен з перерахованих підходів розглянуто як гіпотезу. При роботі системи використовуються оцінки статистичних характеристик розподілу термінів, які мають обмежену точність і якісний характер.

У цьому розділі розглянуто модель подання документів “множина слів” – на підставі пар, процес реферування та організація індексної структури. Запропоновано модифікований алгоритм вибору найбільш інформативних речень з тексту документа. Описано особливості реалізації даної моделі та методу в експериментальній системі, обґрунтовано вибір, представлений загальний алгоритм роботи системи. Для моделі з парами слів описаний

алгоритм відбору пар для обробки. Представлено загальну схему розробленої системи та бібліотеку класів.

Для представлення, опису документа була запропонована власна індексна структура, що складається з двох частин:

- 1) словник. Словник містить терміни, кожен термін має покажчик на пост-лист;
- 2) пост-листи. Кожен пост-лист являє собою масив адрес входжень слова. Адресою слова є ідентифікатор документа і позицій слова в документі.

Можна зробити наступний висновок з приводу організації індексної структури: можлива реалізація індексної структури з використанням словника (Dictionary) та B + дерева. При цьому можна ефективно реалізувати функції, необхідні для зважування документів при пошуку.

Представлено схему розробленої інформаційно-пошукової системи (рисунок 2) та алгоритм обробки запиту.

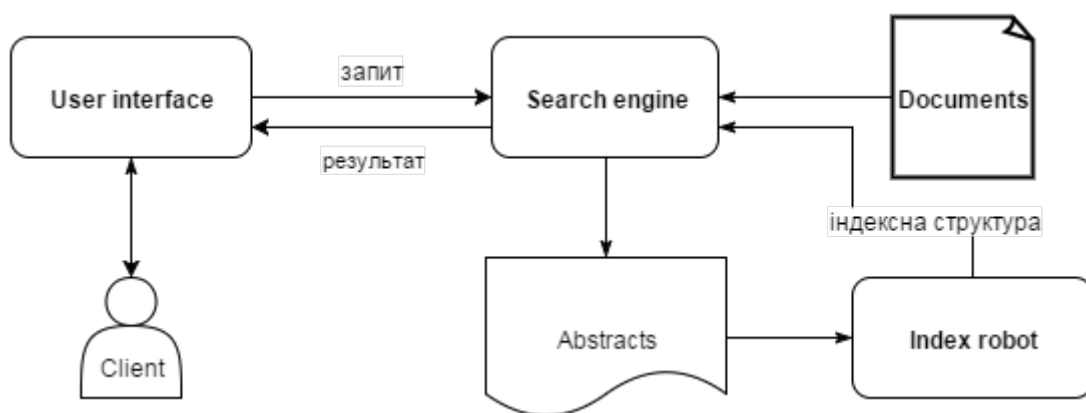


Рисунок 2 – Схема інформаційно-пошукової системи

У шостому розділі представлено результати експериментів, виконаних з метою перевірки ефективності методів і алгоритмів, описаних в попередніх розділах цієї роботи. Для кожної групи експериментів описуються колекції документів, на яких виконувалися експерименти, постановка експерименту і отримані результати.

Із описаних експериментів можна зробити наступні висновки:

- використання контактних пар з невеликою відстанню між словами для інформаційного пошуку призводить до зниження якості інформаційного пошуку. Причиною цього є занадто сильне обмеження, яке призводить до зменшення кількості релевантних документів у видачі, тим самим знижується повнота і якість пошуку. Експериментальні дані показують, що для того, щоб не відбувалося погіршення повноти, потрібно збільшити відстань між словами при формуванні пар. Стійке збільшення якості інформаційного пошуку досягається тільки при використанні досить великої відстані між словами – 10-15 слів. При такій відстані кількість пар слів стає неприйнятно великим для побудови індексу, що містить всі пари;

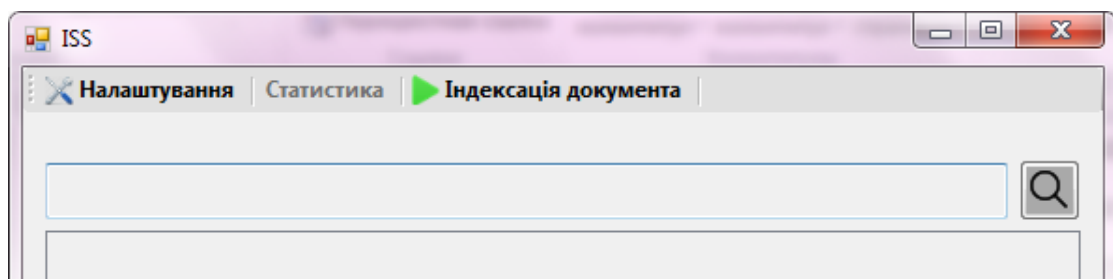
- неможливість використання спеціального індексу робить пошук по парам набагато менш привабливим, тому що при скануванні або зберіганні в індексі повної координатної інформації можна використовувати значно складніші алгоритми виділення і обробки словосполучень;

- попереднє автоматичне реферування документів за допомогою представленої моделі та методу зменшує витрати у перерахунку на один документ і дозволяє значно скоротити час пошуку документів.

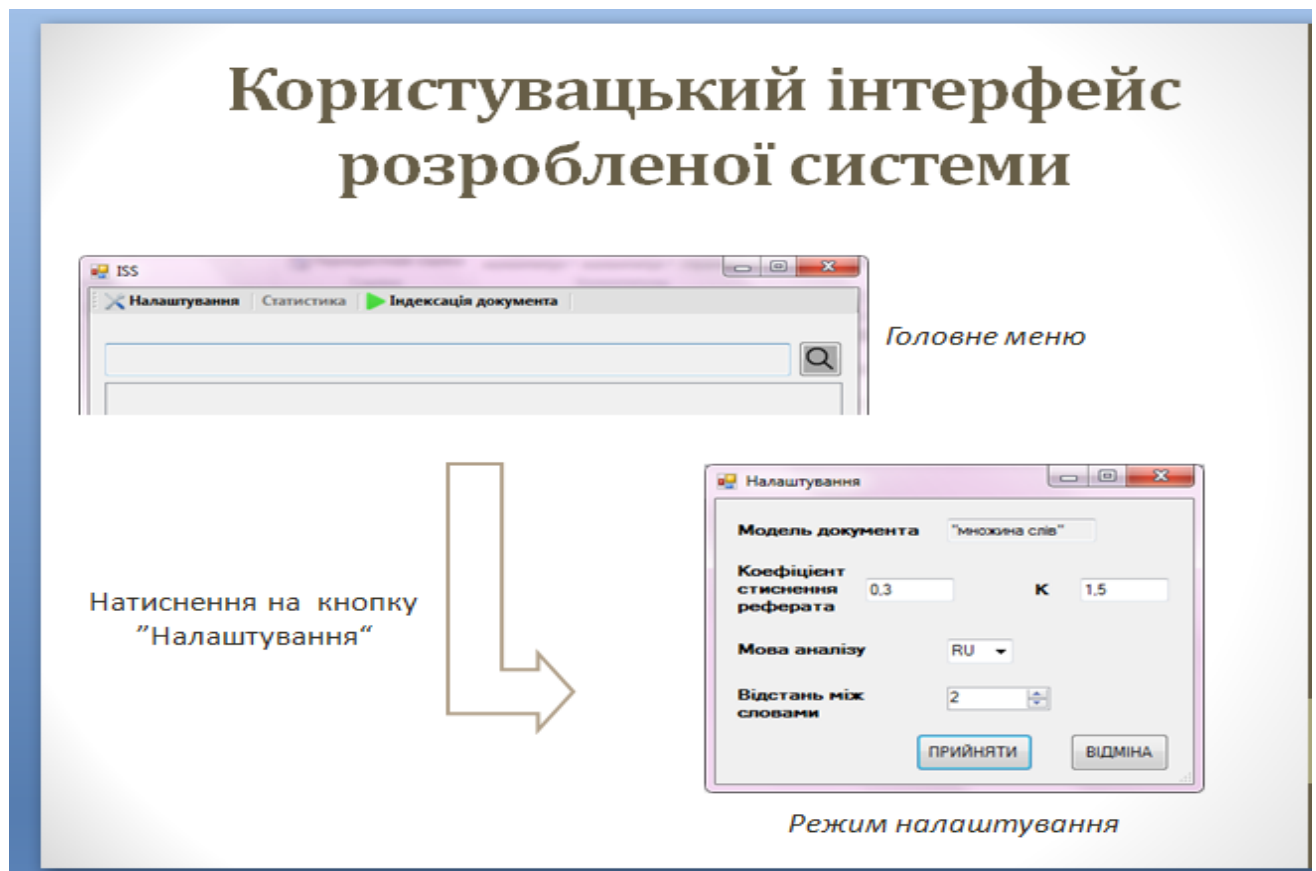
У роботі створено DeskTop застосування та користувацькі засоби для технологічного проведення дослідження запропонованої моделі та методу представлення документа. Зразки роботи користувача у системі наведені на наступних рисунках.

Користувачу потрібно виконати наступну послідовальність дій

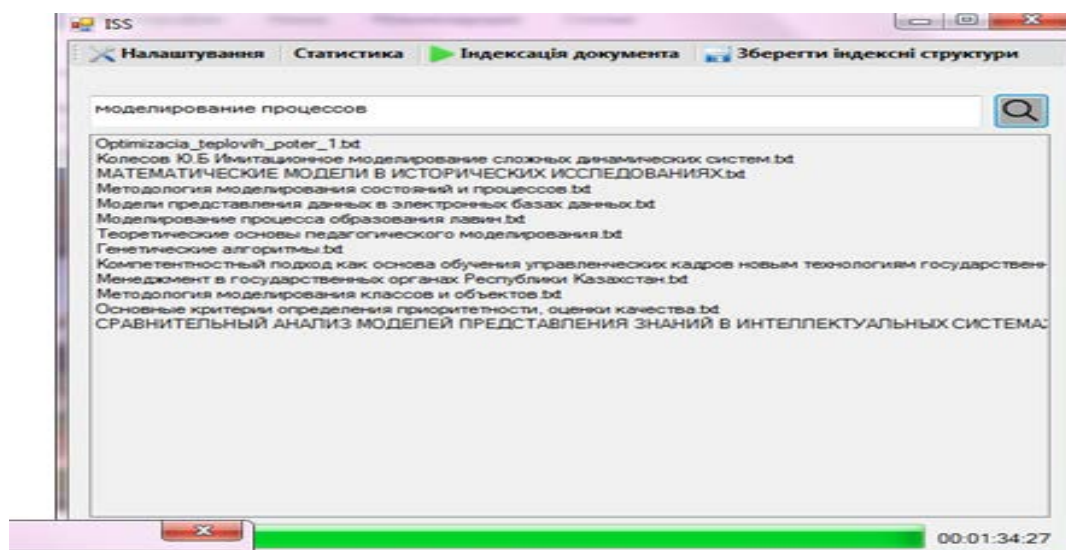
- Потрібно зафіксувати дані запиту у полі ПОШУКУ на головній формі застосування та створити індексну документа (путь доступу до файла)



- Натиснути кнопку НАЛАШТУВАННЯ для завдання параметрів алгоритму створення моделі документа



- натиснути кнопку СТАТИСТИКА
- отримати реферат документу
- зберегти індексну структуру документа



ВИСНОВКИ

В роботі отримані наступні основні результати:

1) розглянуто існуючі методи представлення текстового документа в системах інформаційного пошуку: документ, як множина ваг термінів, документ, як множина фрагментів, документ, як вузол гіпертекстового графа;

2) розроблений варіант реалізації функції зважування документів при інформаційному пошуку, що враховує взаємне положення слів, який дозволив значно збільшити якість інформаційного пошуку;

3) запропонована оригінальна реалізація індексних структур, що дозволяє ефективно виконувати інформаційний пошук з використанням описаних моделей, що враховують взаємне положення слів;

4) реалізований прототип системи, що здійснює інформаційний пошук з використанням описуваних моделей документів;

5) проведена серія експериментів по оцінці якості пошуку системи з використанням власних методик;

6) показано, що використання при інформаційному пошуку моделей документа, що враховують взаємне положення слів, покращує якість пошуку;

7) показано, що пошук з використанням запропонованих моделей документів може бути ефективно реалізований тільки з використанням спеціальних індексних структур;

8) удосконалено модель та метод представлення текстового документа за рахунок попереднього реферування документа та генерування власної індексної структури, зручної у використанні системи під час пошуку для інформаційної.

Деякі моделі неможливо реалізувати в повному обсязі, тому важко оцінити їх вплив на якість пошуку. Інші поки що недостатньо вивчені з точки зору їх впливу на якість пошуку і вимагають додаткових досліджень.

Розглянута модель “множини слів” забезпечує якісне виконання пошукових операцій, тому що більшість пошукових запитів користувачів містять одне-два слова, що для алгоритмів виділення словосполучень є достатньо легко

реалізуємим. Крім всього, статистика розподілу пар слів дуже схожа на статистику щодо рідко зустрічаючих в колекції термінів і є порукою у використанні добре досліджених і надійних алгоритмів обробки термінів для моделей, які не враховують взаємне положення слів.

У порівнянні з системами реферування, в яких проводиться повний цикл, обробки документів, даний метод дозволяє значно скоротити тимчасові витрати на складання реферату.

Наукові результати магістерської дисертації набули подальшого розвитку застосування моделей та методів представлення текстового документа в системах інформаційного пошуку для текстових колекцій та оцінка якості пошуку.

Запропонований у роботі метод пошуку може бути з успіхом застосований у сучасних інформаційних та інформаційно-аналітичних системах для оброблення великих масивів документів. Попереднє автоматичне реферування документів зменшує витрати у перерахунку на один документ і дозволяє значно скоротити час пошуку документів.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Пляшко Н. В. Моделі і методи представлення текстового документа в системах інформаційного пошуку / Н. В. Пляшко, О. О. Гагарін // Тези доповідей XIV Міжнародної науково-практичної конференції аспірантів, аспірантів і студентів, присвяченої 85 річчю теплоенергетичного факультету “Сучасні проблеми наукового забезпечення енергетики”, 18-21 квітня 2016 р., м. Київ. – К.: НТУУ “КПІ”, 2014. – 2 т. – С. 117.