НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
"КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

Моделі і методи представлення текстового документа в системах інформаційного пошуку

Виконала
студентка 6 курсу ТЕФ групи ТІ-41м
Пляшко Н.В.
Керівник проекту
к.т.н., доц. Гагарін О.О.

Київ — 2016

Мета дослідження

- Удосконалення моделей і методів представлення текстового документа в системах інформаційного пошуку, які враховують взаємне розташування слів в документі;
- Оцінити їх вплив на якість інформаційного пошуку ;
- Створити засоби їх ефективної реалізації з використанням індексних структур та попереднього автоматичного реферування документів.

Об’єктом дослідження є програмне забезпечення систем інформаційного пошуку.

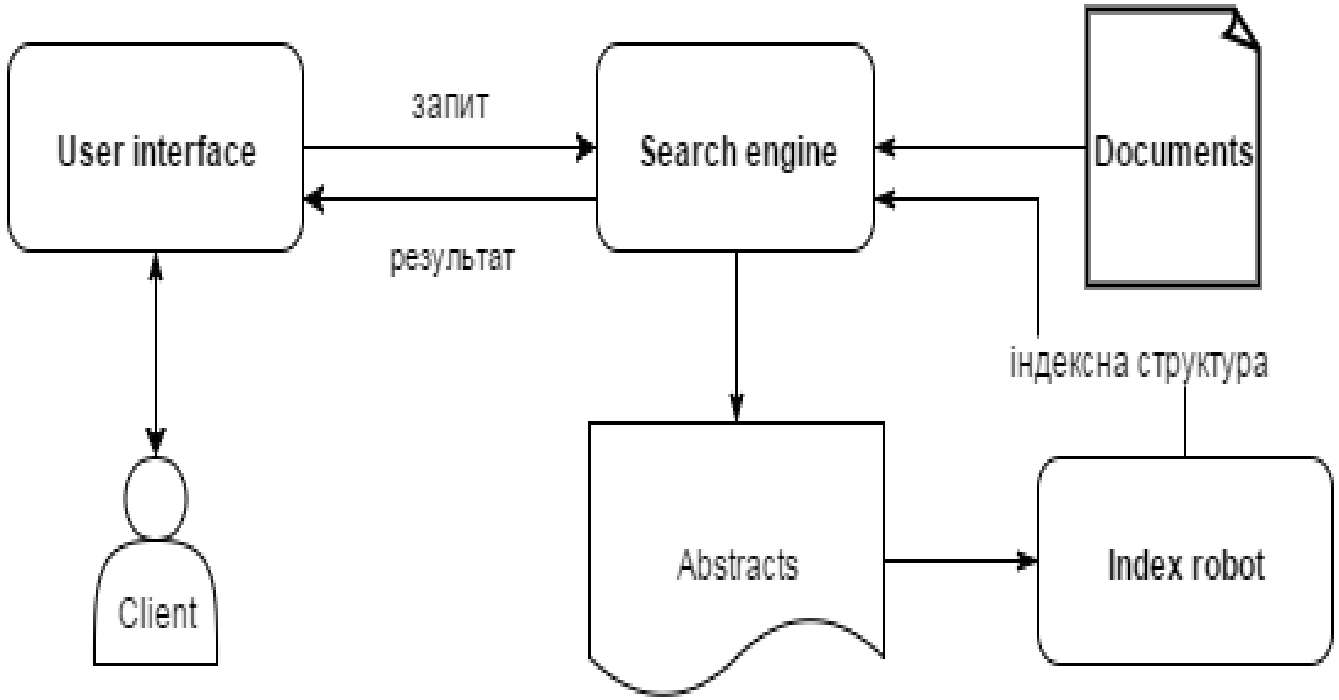
Предмет дослідження – програмне забезпечення для реалізації моделей та методів представлення текстових документів в системах інформаційного пошуку.

Завдання дослідження

- проаналізувати існуючі моделі і методи інформаційного пошуку
- дослідити вплив різних характеристик моделі на якість інформаційного пошуку
- обґрунтувати доцільність використання попереднього реферування документів у процесі пошуку
- удосконалити модель і метод представлення текстового документа
- розробити програмне забезпечення реалізації моделі та обраного методу представлення текстового документа для операцій пошуку

Загальна схема процесу пошуку

Схема роботи інформаційно-пошукової системи



Класифікація методів інформаційного пошуку

- Класичний інформаційний пошук
- Автоматична кластеризація, рубрикація або фільтрація документів
- Виділення інформації з тексту (text mining)

Проблеми оцінки якості інформаційного пошуку

- вибір показників оцінки пошуку
- вплив тестового набору даних
- відмінності в реалізації компонентів системи інформаційного пошуку
- висока обчислювальна вартість

Критерії ефективності пошуку

- Поняття релевантності, тобто відповідності документа запиту
- Повнота пошуку – здатність ІПС видавати всі релевантні документи
- Точність пошуку – здатність ІПС відсіювати нерелевантні документи
- Зусилля, що витрачаються на формулювання запитів, взаємодія з системою і перегляд виданої інформації
- Форма представлення знайденої інформації
- Повнота інформаційного масиву, тобто ступінь охоплення всіх релевантних інформаційних ресурсів, що цікавлять користувачів

Моделі інформаційного пошуку

- Документ, як множина слів (термінів)
  - Бінарна модель
  - Модель ваг слів
- Документ, як множина фрагментів
- Документ, як вузол гіпертекстового графа

Модель “множини слів”
Проблеми виділення слів

- Особливості алфавітів і мов
- Кодування текстів
- Різні формати подання документів
  - застарілі і закриті комерційні формати
  - формати, орієнтовані на графічне представлення документа
  - формати, які містять динамічно виконуючі елементи
- Морфологічний аналіз (стемінг)

Модель “множини слів”

Підходи визначення “ваги” слова

- Статистичний підхід визначення “ваги” слів
- Місце появи слів
- Форматування слова

Моделі визначення “ваги” (зважування термінів)

- Частотна модель
- Імовірнісна модель
- Латентно-семантичний аналіз

Модель “множини слів”

Статистичні методи визначення “ваги” (зважування слів)

- кількість появи слова в документі (TF)
- частота появи слова в документі $TF * IDF$

  Для довгих документів застосовується доповнена нормалізована частота $0.5 + 0.5(TF / ATF)$,

  де ATF – середня частота терміна в документі
- логарифм частоти входження слова $1 + \log(TF)$

  Для стійкості до переоцінки документів застосовується формула $(1 + \log(TF)) / (\log(MTF))$,

  де MTF – максимальна частота слова в документі

Модель “множини фрагментів”
Підходи розбиття документа на фрагменти

- *Passage retrieval*. Документ розбивається на множину фрагментів, які розглядаються як об’єкти, по яким здійснюється пошук. З точки зору моделі документа, він являє собою не одну множину термінів, а кілька пов’язаних між собою множин
- *Поділ на фрагменти однакового по розміру об’єму*. При цьому розподіл є статичним і проводиться на етапі індексування
- *“Ковзаюче вікно”. Для* кожного запиту вибирається своє, найбільш підходяще для нього розбиття. Враховуються фрагменти, які отримали найбільшу вагу або вага документа являє собою суму значень всіх вікон

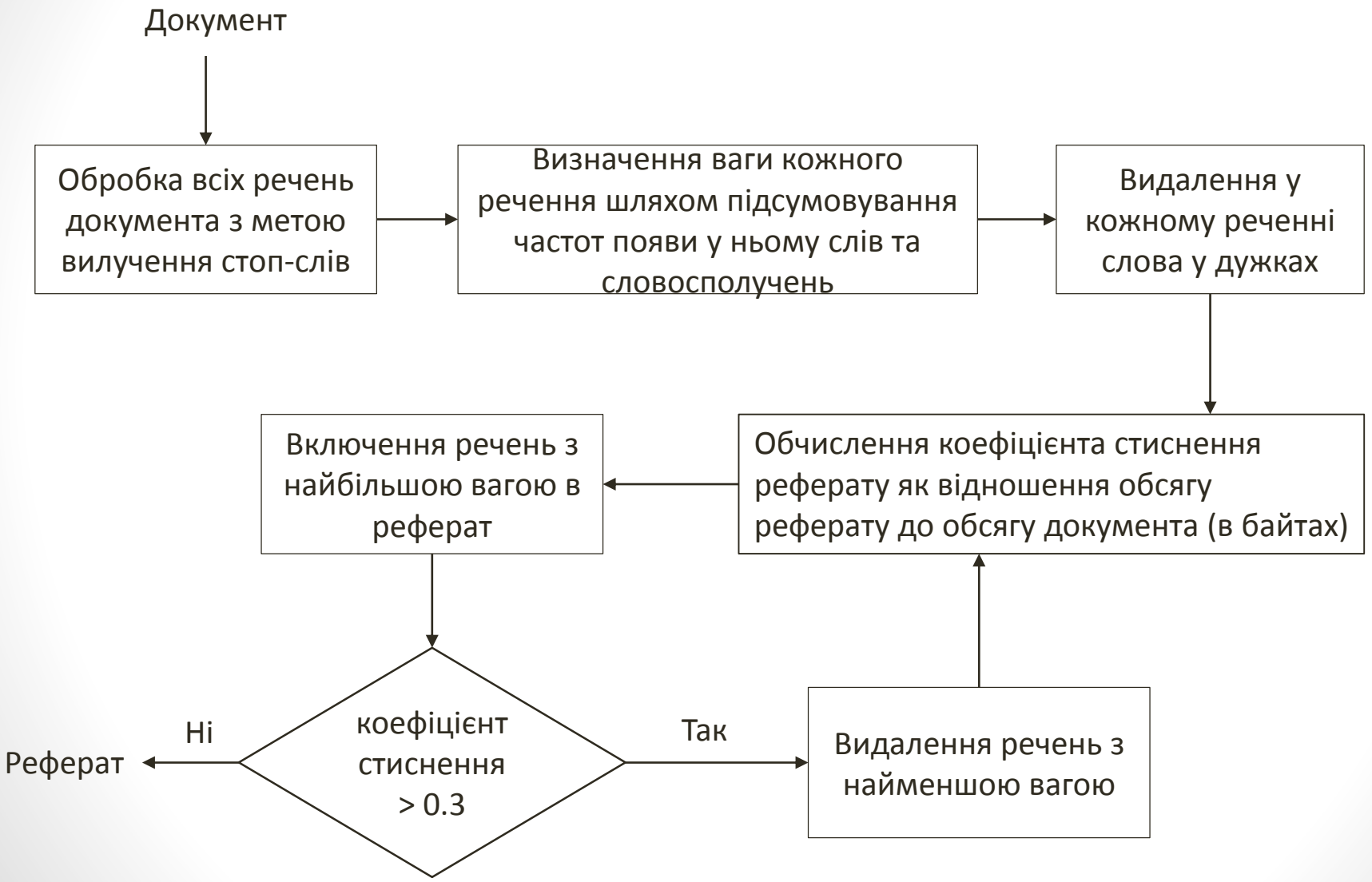
Модель “багато зв’язного графа”

Гіпертекстові посилання між документами

- Модель, заснована на “Індексному цитуванні” або PageRank
- Модель, що враховує контекст запиту
- Модель, що “переносить” терміни

Узагальнена схема роботи системи

Схема процесу реферування



Процес індексування

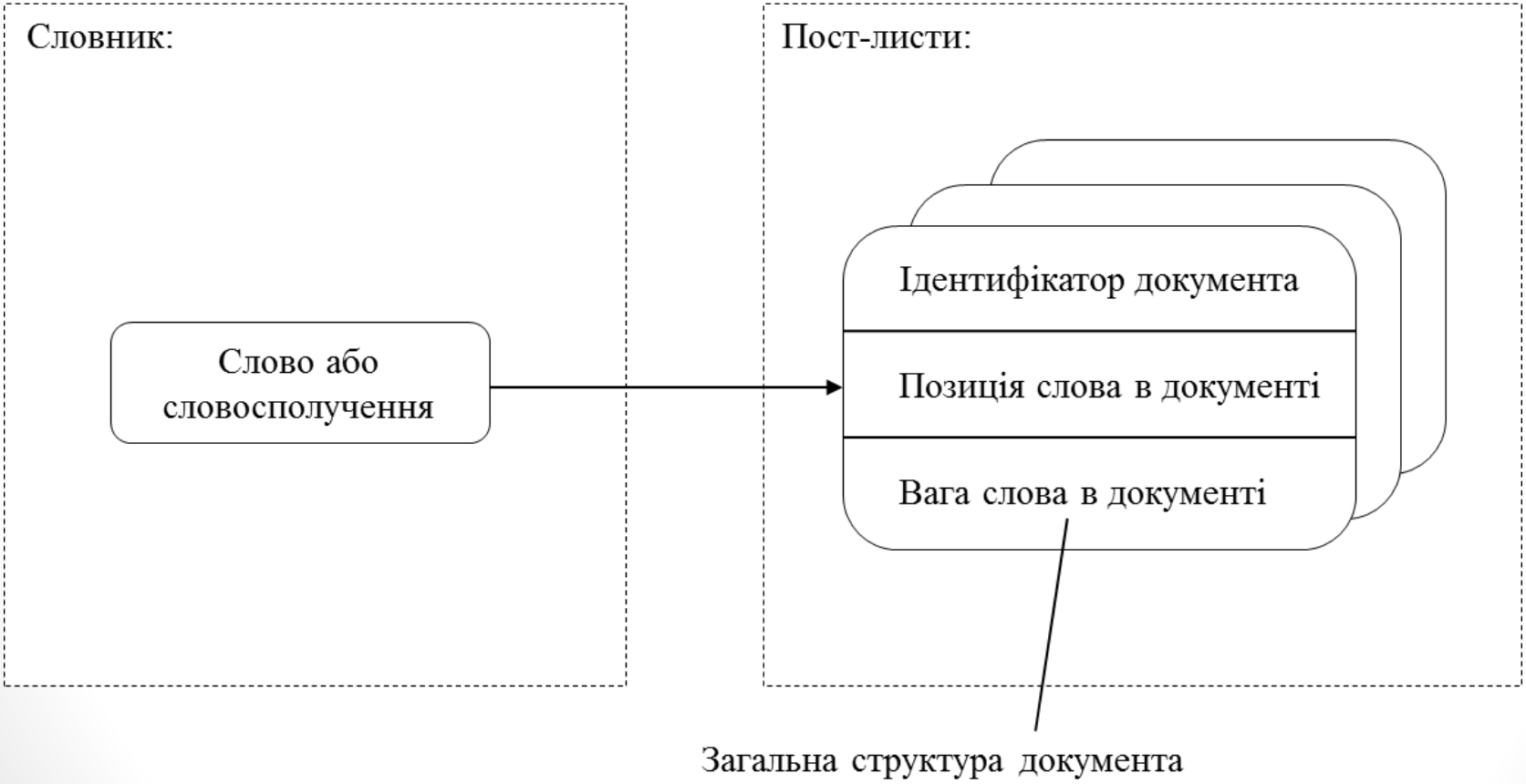
Розбиття реферату на множину термінів

Приведення всіх слів до словарної форми та процес стемінгу. Слова із латинських букв, чисел, стоп-слів не враховуються

Відбір індексаційних термінів рефератів, які використовуються для опису змісту документів

Приписування термінам деякої ваги, яка відображає передбачувану важливість термінів

Схема індексної структури документа



Загальний алгоритм обробки запиту розробленою системою

- Попереднє реферування документа
- Індексування реферату

  Базова вага слова обчислюється за формулою $TF * IDF$

  де TF (частота слова) – відношення числа входження деякого слова до загальної кількості слів документа,

  IDF (зворотна частота документа) – інверсія частоти, з якою деяке слово зустрічається в документах колекції:

  $$IDF = \log \frac{|D|}{|(d_i \supset t_i)|}$$

  де $|D|$ — кількість документів колекції

  $|(d_i \supset t_i)|$ — кількість документів, в яких зустрічається слово

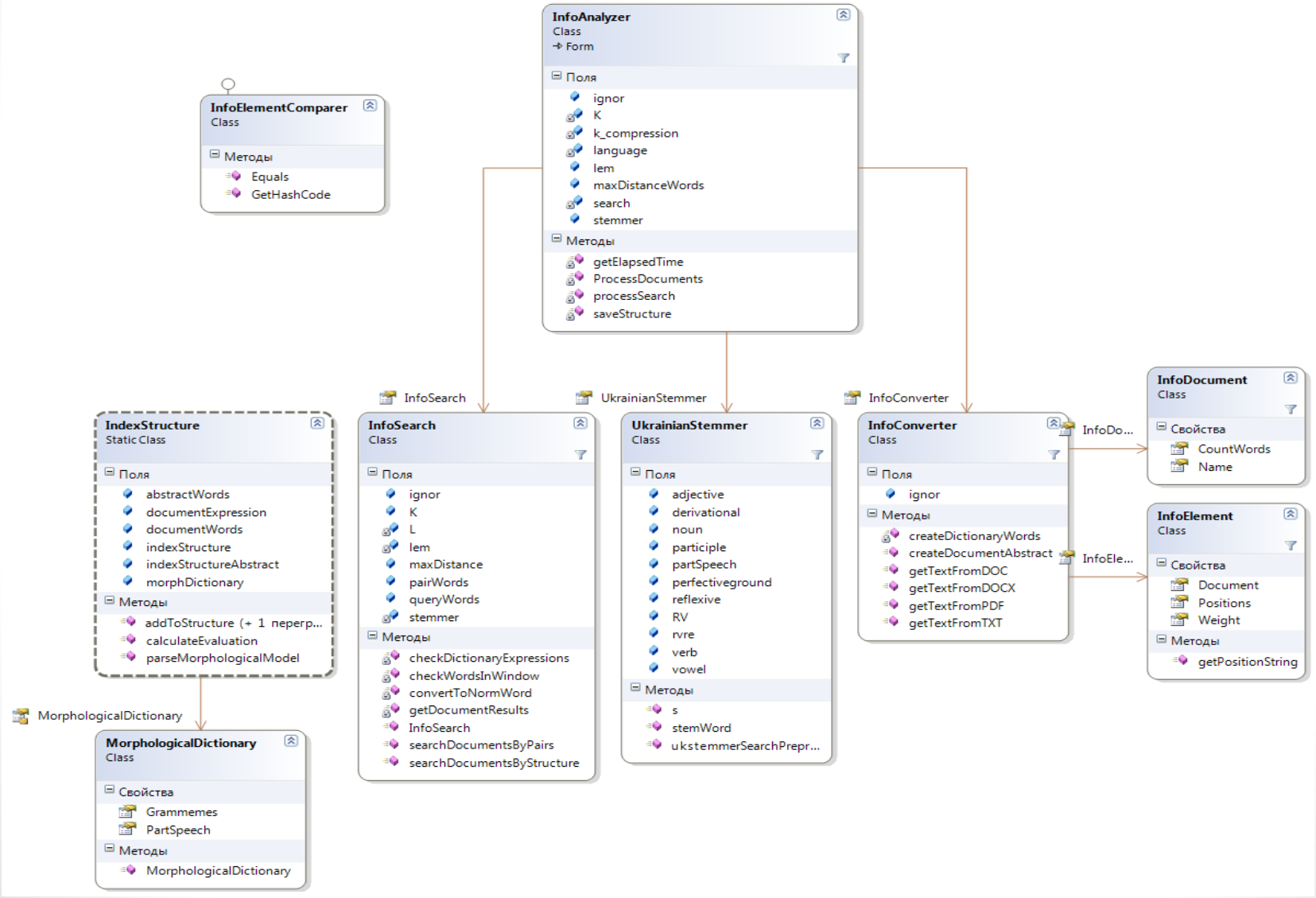
- Пошук, результати пошуку виводиться у вигляді списку документів, упорядкованих по спаданні ваги

Алгоритм визначення стійкості словосполучення

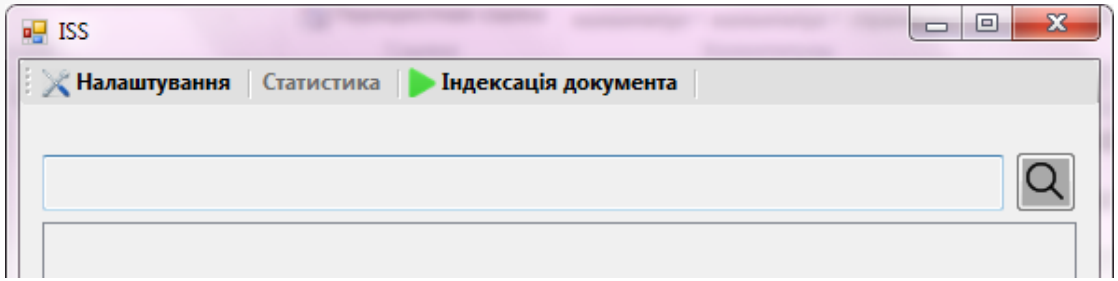
1) для кожної пари виходить список документів, які містять цю пару S_{pair}.
2) для кожного слова, що входить в пару, також формуються списки входження s_{w1} і s_{w2}
3) формується перетин списків для слів, що входить в пару – $S_c = S_{w1} \cap S_{w2}$

4) якщо $|S_{pair}| * K > |S_c|$, то дана пара вважається словосполученням і залишається, інакше відкидається

Діаграма класів розробленої системи

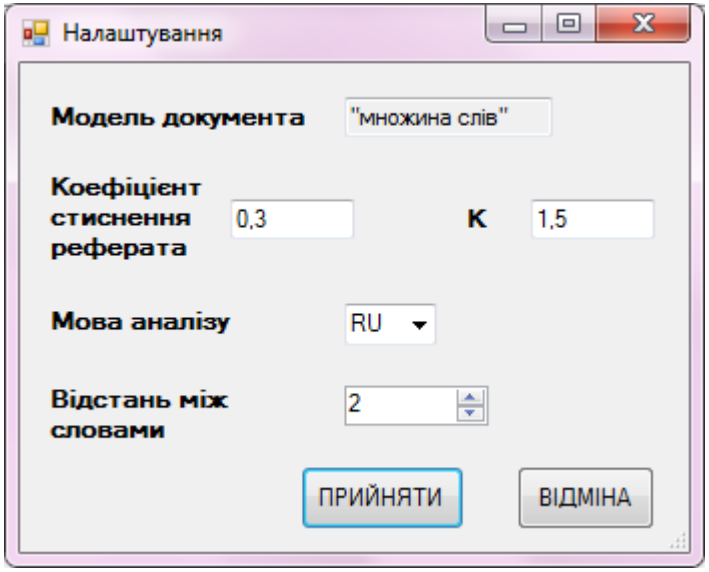


Користувацький інтерфейс розробленої системи



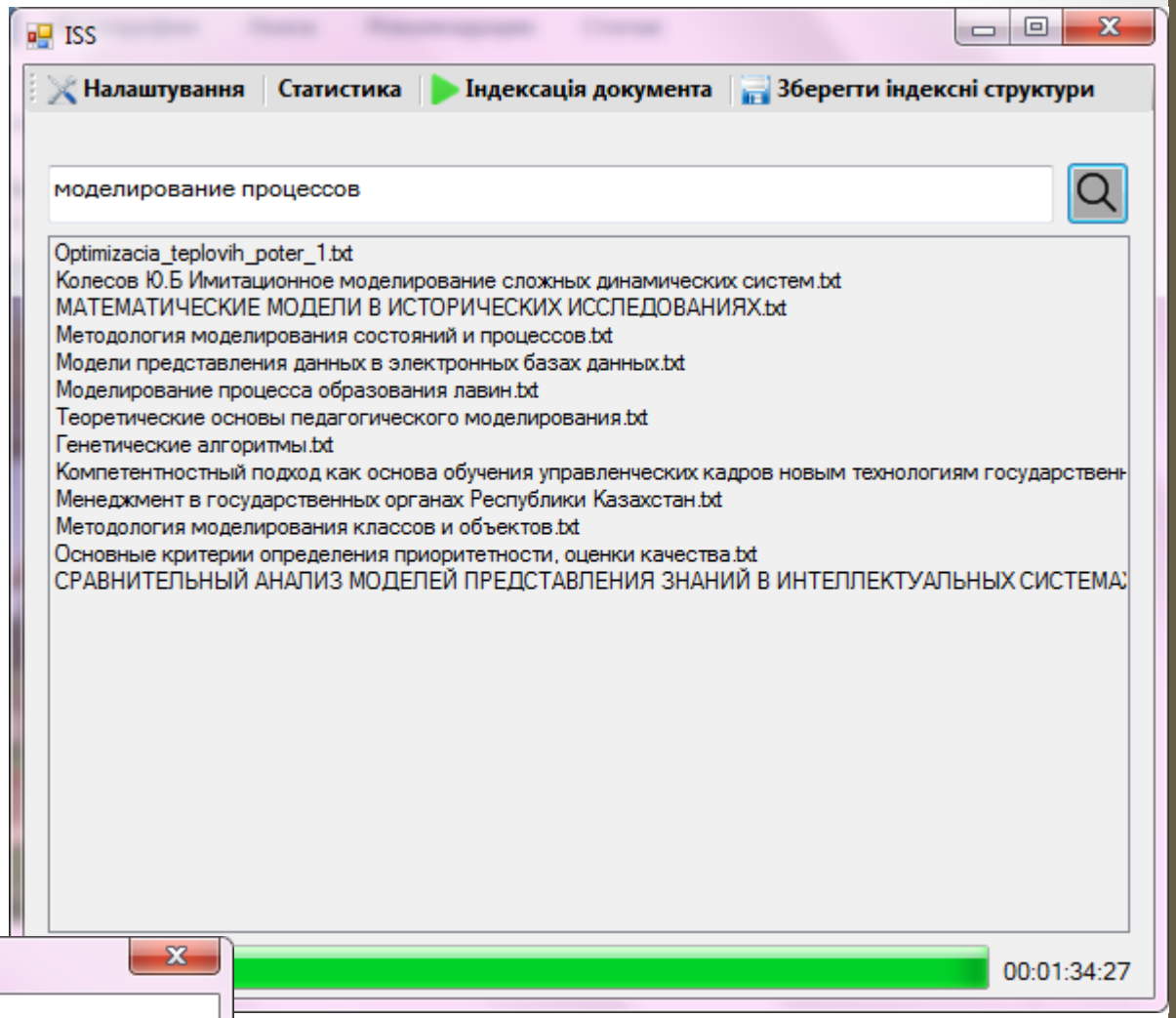
Головне меню

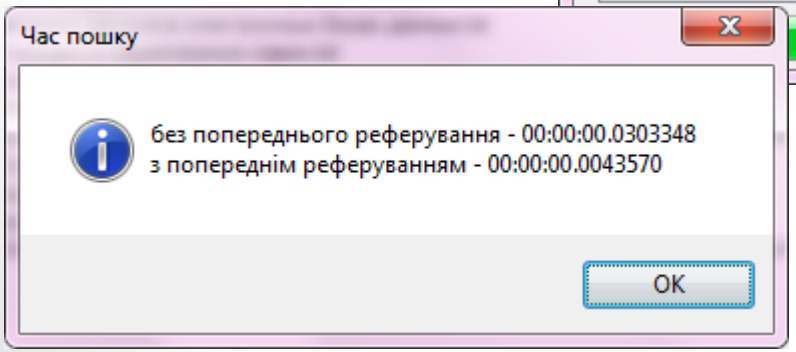
Натиснення на кнопку ”Налаштування“



Режим налаштування

Натиснення на кнопку ”Статистика“





Наукова новизна

- удосконалено метод представлення текстового документа за рахунок введення процедури попереднього реферування документа та генерування власної індексної структури, що спрощує та прискорює роботу системи пошуку
- удосконалено модель зважених термінів за рахунок введення додаткових характеристик ваг та введення механізму оцінки стійкості словосполучень, що підвищує ефективність результатів пошуку
- набуло подальшого розвитку виявлення критеріїв оцінювання якості результатів пошуку за рахунок введення текстових колекцій та сортування за ознаками вагових показників реферату.

Апробація результатів

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.)

Висновки

- запропонована оригінальна реалізація індексних структур, що дозволяє ефективно виконувати інформаційний пошук з використанням описаних моделей, що враховують взаємне положення слів
- реалізований прототип системи, що здійснює інформаційний пошук з використанням описуваних моделей документів
- проведена серія експериментів по оцінці якості пошуку системи з використанням власних методик
- показано, що використання при інформаційному пошуку моделей документа, що враховують взаємне положення слів, покращує якість пошуку
- удосконалено модель та метод представлення текстового документа за рахунок попереднього реферування документа та генерування власної індексної структури, зручної у використанні системи під час пошуку

Дякую за увагу