Київський національний економічний університет імені Вадима Гетьмана

Аналіз методів автоматизованого пошуку електронних документів в великих слабо структурованих масивах

Виконала: студентка
групи 1 заочної форми навчання
факультету «»
Хххxx хххххх хххх

Метою дослідження є виконання аналізу методів та технологій і визначення необхідності автоматизованого пошуку електронних документів в великих слабоструктурованих масивах.

Об’єктом дослідження є інформаційні технології виділення та обробки знань.

Предметом дослідження – технологія Text Mining для автоматизованого пошуку електронних документів у великих слабо структурованих масивах.

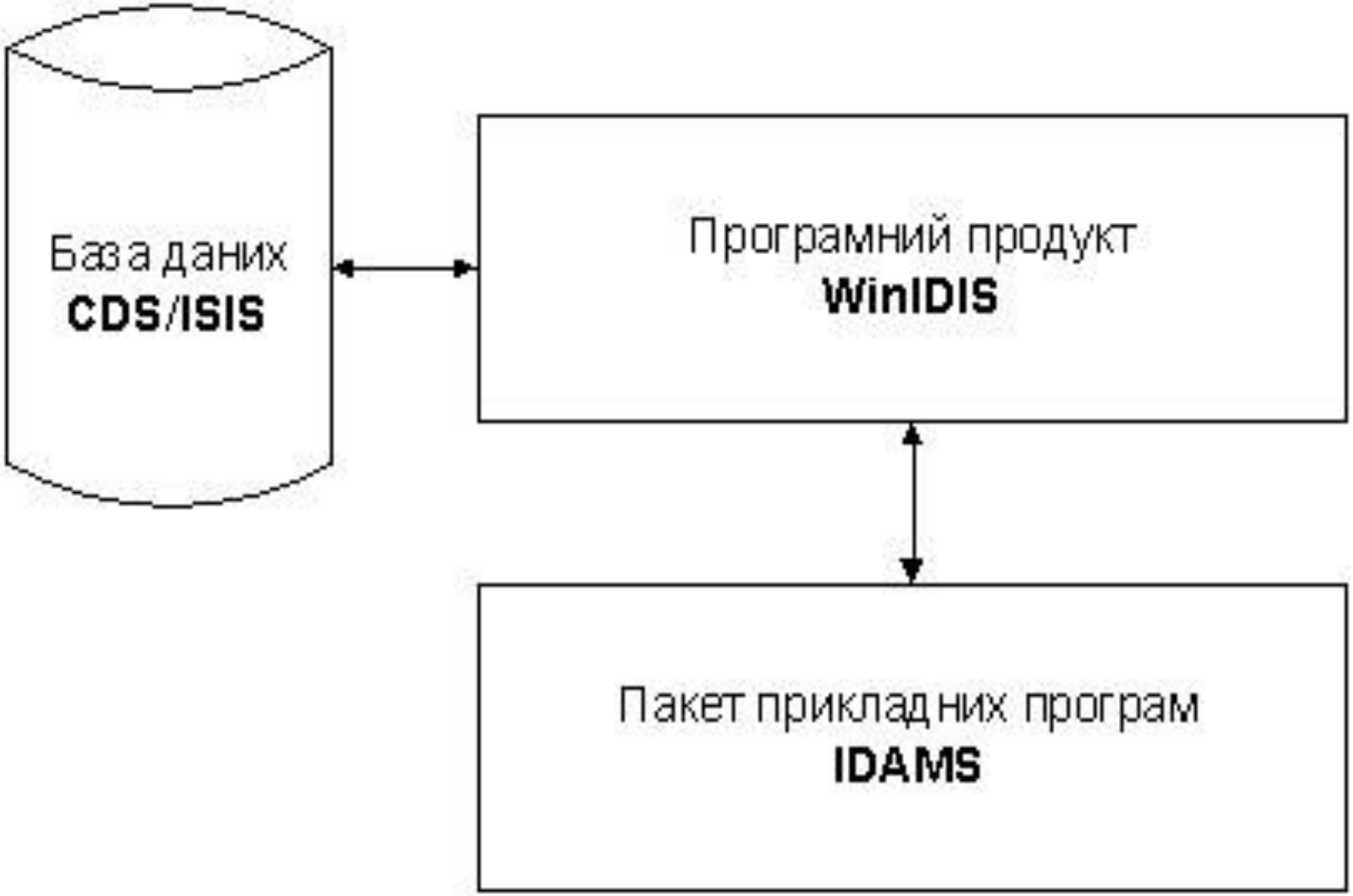
Значення інформаційних ресурсів

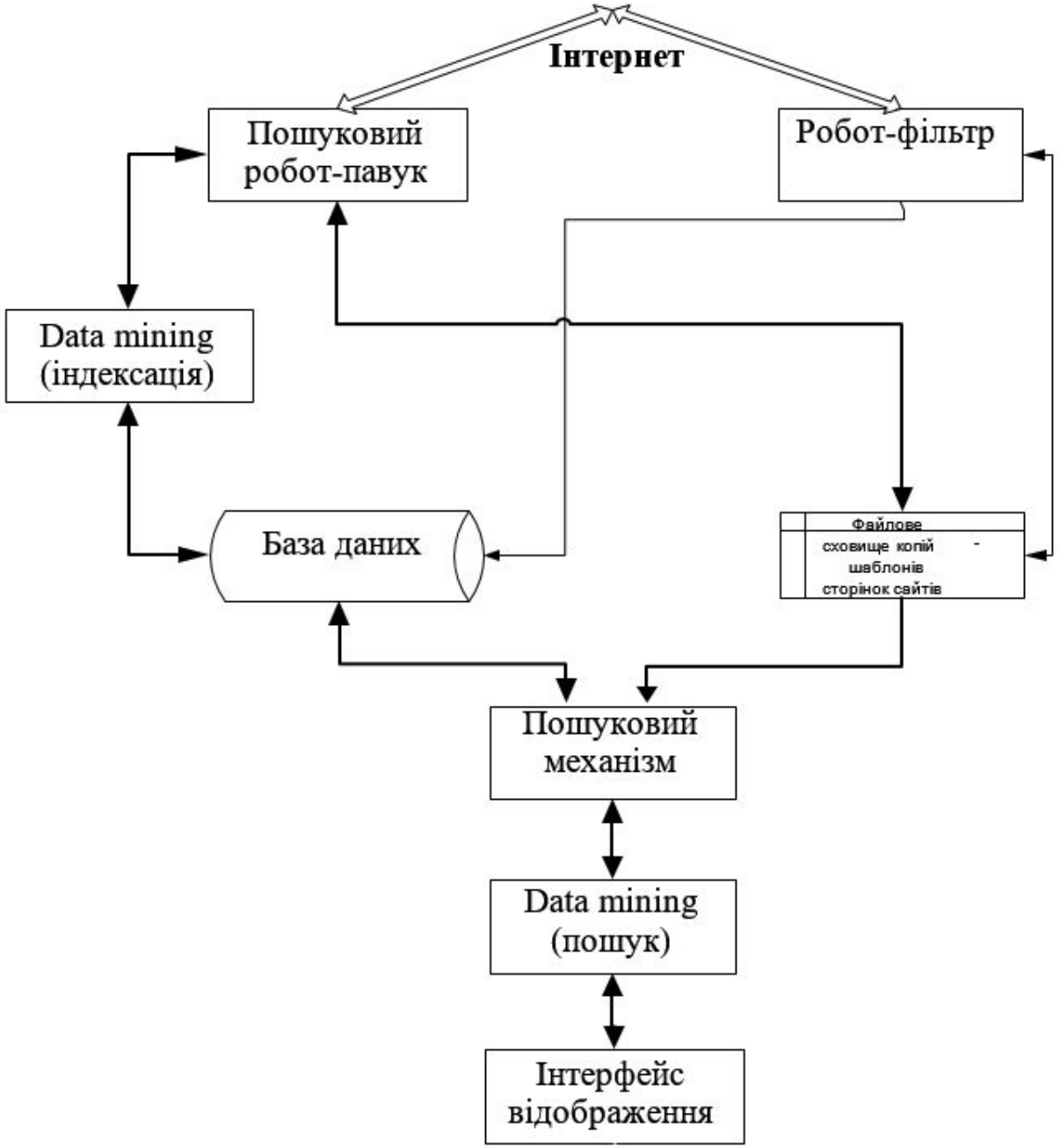
Інформаційні технології виділення та обробки знань

Інформаційна технологія базується на реалізації інформаційних процесів, різноманітність яких вимагає виділення базових, характерних для будь-якої інформаційної технології. Базовий технологічний процес заснований на застосуванні стандартних моделей і інструментальних засобів. Серед **базових технологічних процесів** виділяють:

- отримання інформації;
- транспортування інформації;
- обробку інформації;
- зберігання інформації;
- представлення та використання інформації.

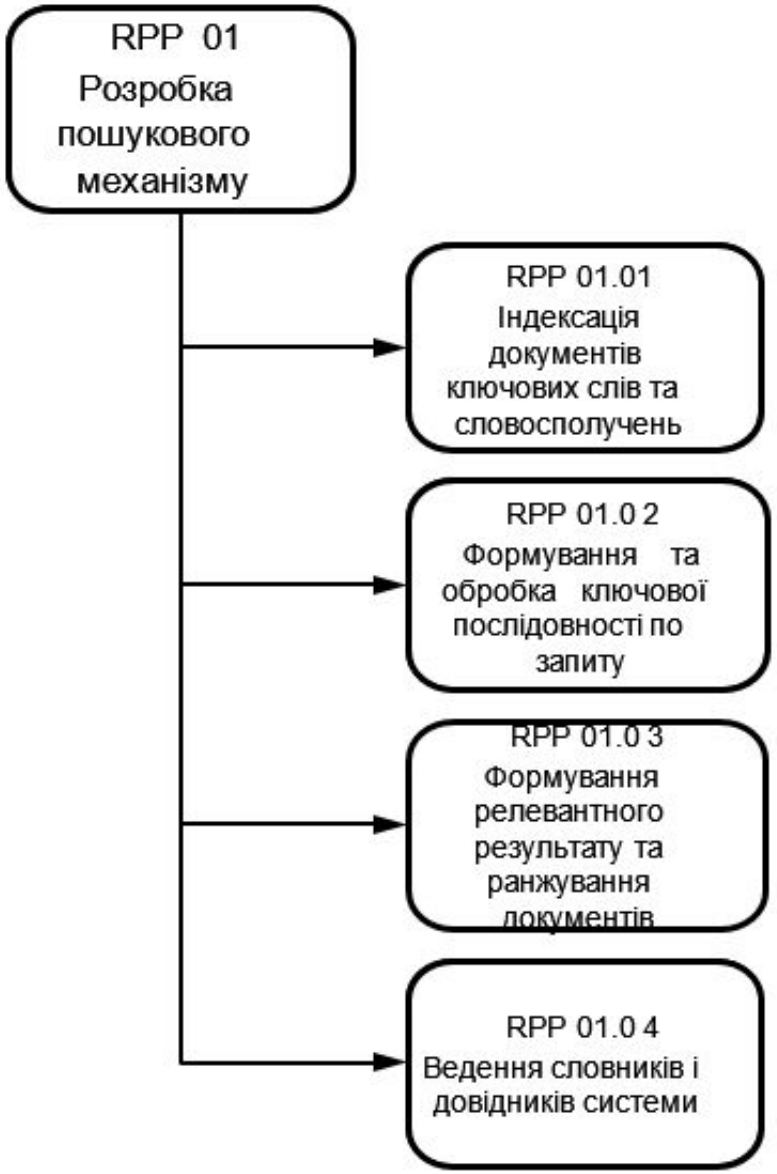
Аналіз програмного забезпечення для виявлення текстових документів





Загальна схема роботи системи пошуку і аналізу тексту

Функціонування механізму роботи пошукової системи можна поділити на два основні, незалежні один від одного завдання: індексація метаданих отриманих від пошукового агента і організація пошуку на підставі запиту користувача і індексованих в системі документів.



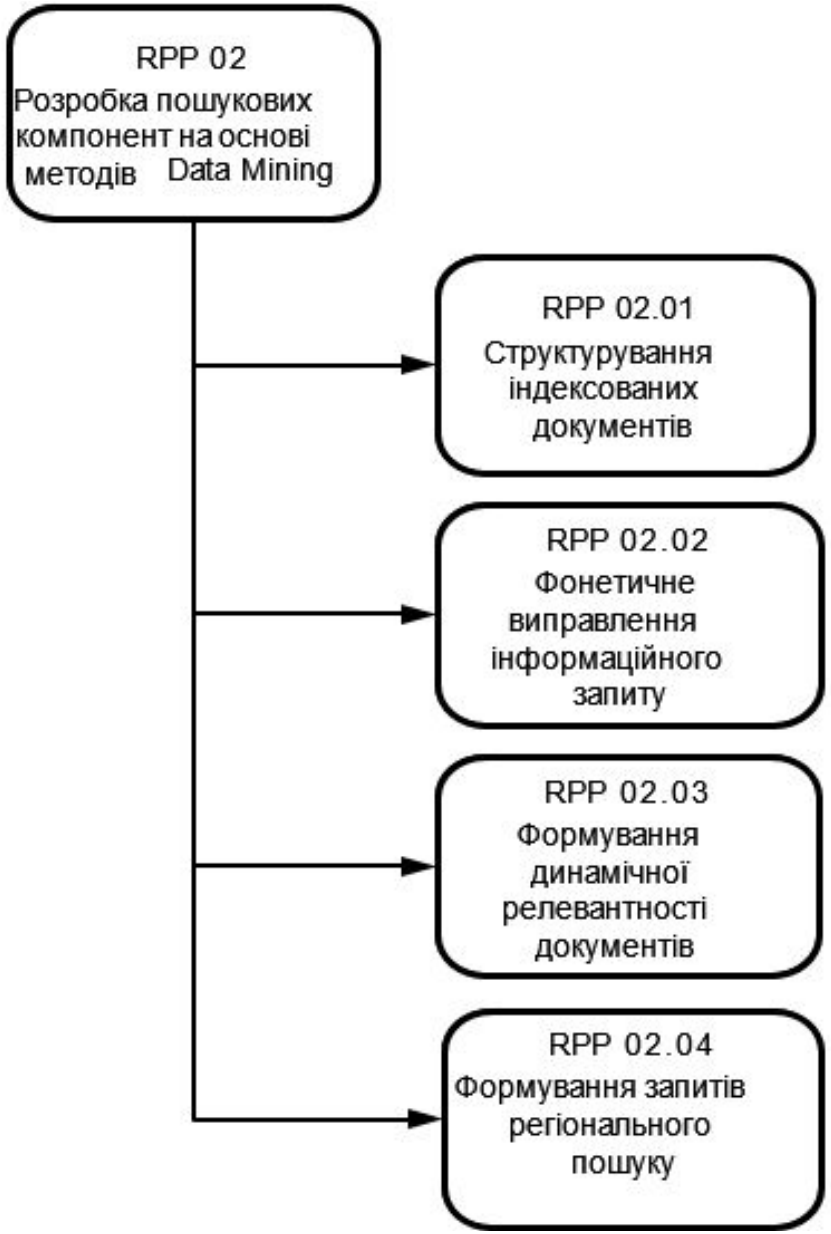
Діаграма дерева функцій структурування тексту під час пошуку

Функція «Індексація документів, ключових слів і словосполучень» призначена для перетворення отриманої від «агента» інформації у оптимальний для системи вигляд і додавання її до бази даних.

Функція «Формування релевантного результату та ранжування документів» забезпечує організацію ранжування документів по мірі відповідності до запиту в залежності від наявності метаданих у термінах і їх частоти.

Функція «Ведення словників пошуку» призначена для організації роботи зі словниками термінів, сто-слів, атрибутів та ін. необхідних для прискорення процесу пошуку.

Функція «Формування ключової послідовності по запиту» необхідна для перетворення запиту користувача у прийнятний для системи вигляд.



Діаграма дерева функцій пошукових компонент на основі методів Data Mining

Функція «Структурування індексованих документів» призначена для перетворення отриманої інформації до структурованого вигляду за для забезпечення швидкості обробки текстів, підвищення її якості за допомогою «самонавчання».

Функція «Фонетичне виправлення інформаційного запиту» виконує підвищення релевантності пошукової системи шляхом корегування помилкових інформаційних запитів.

Функція «Формування динамічної релевантності документів» виконує аналіз callback'ів та на основі отриманих відомостей підвищує або понижує релевантність документа.

Функція «Формування запитів регіонального пошуку» забезпечує організацію пошуку інформації по вибраним державним регіонам.

Методи ранжирування на основі машинного навчання

Позначимо запит користувача буквою q, а документ - буквою d. Метод зваженого зонного ранжирування присвоює парі (q, d) значення релевантності на відрізку [0..1], обчислюючи лінійну комбінацію зонних показників, до якої кожна зона документа вносить булеве значення. Розглянемо безліч документів, кожен з яких має l-зон. Нехай g1, g2..gi $\in$[0.1], так що:

$$\sum_{i=1}^{l} g_i = 1$$

Нехай Si де 1<i < L, - булева величина, що означає відповідність(або її відсутність) між запитом q і i-й зоною. Це відображення може здійснювати будь-яка булева функція, що відображає наявність термінів запиту в зоні в множині {0, 1}. Таким чином, зважену зонну релевантність можна визначити за ф$\sum_{i=1}^{l} g_i * s$о:

Ваги g1 .. gi вказуються експертами або користувачем. Однак набагато частіше ваги визначаються на основі навчальних прикладів, оцінених заздалегідь.

Схема відповідностей літер до алгоритму Soundex

Літери	Значення
B, P	1
F V	2
C K S	3
G, J	4
Q, X, Z	5
D, T	6
L	7
M, N	8
R	9

За основу алгоритму Daitch-Makotoff у взято оригінальний Soundex, але він має значно більш складні правила конверсії - тепер у формуванні результуючого коду беруть участь не тільки одиночні символи, а й послідовності з декількох символів.

Крім того, одна комбінація результату забезпечує близько 600 тисяч різних варіацій коду, що у поєднанні з ускладненими правилами зменшує кількість хибнопозитивних термінів у результуючій множині

Схема відповідностей літер до алгоритму Daitch-Makotoff

Початкові буквосполучення	На початку слова	Після голосної	Інші
AI, AJ, AY, EI, EY, EJ, OI, OJ, OY, UI, UJ, UY	0	1	
AU	0	7	
IA, IE, IO, IU	1		
EU	1	1	
A, UE, E, I, O, U, Y	0		
J	1	1	1
SCHTSCH, SCHTSH, SCHTCH, SHTCH, SHCH, SHTSH, STCH, STSCH, STRZ, STRS, STSH, SZCZ, SZCS	2	4	4
SHT, SCHT, SCHD, ST, SZT, SHD, SZD, SD	2	43	43
CSZ, CZS, CS, CZ, DRZ, DRS, DSH, DS, DZH, DZS, DZ, TRZ, TRS, TRCH, TSH, TTSZ, TTZ, TZS, TSZ, SZ, TTCH, TCH, TTSCH, ZSCH, ZHSH, SCH, SH, TTS, TC, TS, TZ, ZH, ZS	4	4	4
SC	2	4	4
DT, D, TH, T	3	3	3
CHS, KS, X	5	54	54
S, Z	4	4	4
CH, CK, C, G, KH, K, Q	5	5	5
MN, NM		66	66
M, N	6	6	6
FB, B, PH, PF, F, P, V, W	7	7	7
H	5	5	
L	8	8	8
R	9	9	9

Технологія аналізу тексту Text Mining містить 4 основні етапи

Перспективи використання Text Mining

В даний час пропонується досить багато інструментів текстомайнінга – від відносно простих програм, що спираються на статистичний аналіз окремих термінів у текстах, таких як WordStat, до найскладніших додатків типу Aerotext і Businessobjects Text Analysis.

З розвитком Інтернету аналіз, що базується на Text Mining, може реалізовуватися не лише за допомогою впроваджуваних в організації додатків, але і у вигляді онлайнового сервісу.

Останнім часом Text Mining аналіз множинних відкритих джерел інформації стає доступним для комерційних, політичних та інших організацій за рахунок появи саме таких онлайнових служб.

Технології видобутку інформації з неструктурованих текстів (Text Mining) використовуються на практиці вже сьогодні, оскільки обсяги доступною і корисною інформацією ростуть з кожним днем, а потреба в їх аналізі є досить актуальною.

Кінець

Дякую за увагу.