Система мультимодального текстового пошуку

Шевченко Тарас Петрович, КА-61м

КПІ ім. Ігоря Сікорського
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

2018

Об’єкт та предмет дослідження

Об’єкт дослідження

- системи текстового пошуку.

Предмет дослідження

- системи мультимодального текстового пошуку.

Мета роботи

Мета роботи

Розробити систему мультимодального текстового пошуку, яка може застосовуватися над даними із попередньо заданою користувачем структурою, при цьому порівняти та вдосконалити існуючі алгоритми текстового пошуку та попередньої обробки даних.

Мультимодальний пошук

Поняття про мультимодальний пошук

Мультимодальний пошук комбінує різні методи текстового пошуку, при цьому може бути застосованим до документів складеної структури.

Існуючі рішення

Проблеми існуючих рішень

1. Не стискають інвертований індекс або стискають методами загального призначення.
2. Використовують схему зважування, у якій ваги можуть бути від'ємними.
3. Відсутність моделі для розширення набору слів пошукового запиту.

Архітектура реалізованої системи

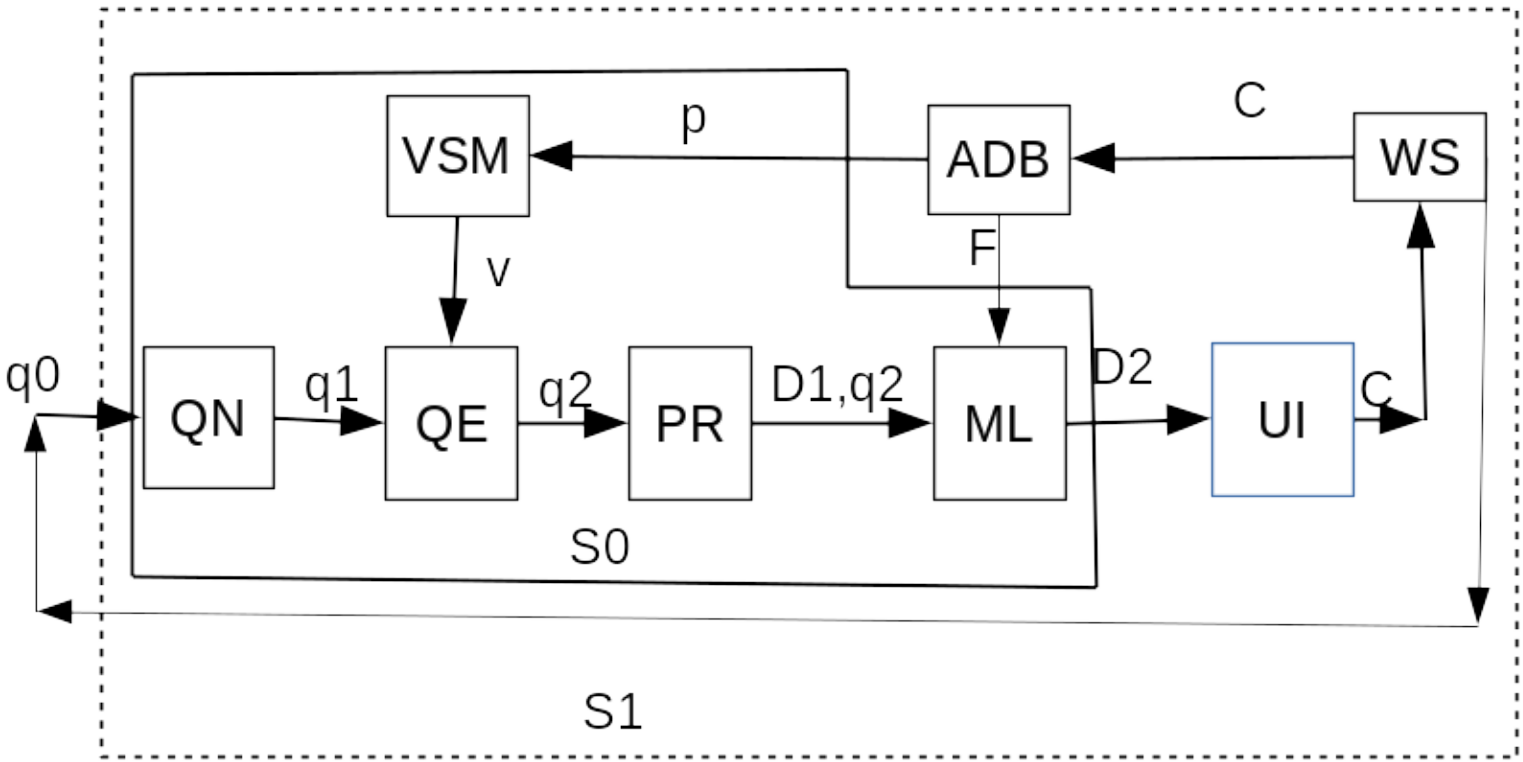

Рис.: Діаграма робити системи мультимодального пошуку

Умовні позначки із опису архітектури системи

1. QN - модуль нормалізації вхідного запиту
2. q_i - пошуковий запит
3. QE - модуль розширення вхідного пошукового запиту
4. PR - ймовірнісна модель текстового пошуку
5. $D1$ - розширений список документів
6. ML - модель машинного навчання,
7. $D2$ - множина релевантних документів
8. UI - користувацький інтерфейс

Умовні позначки із опису архітектури системи (продовдження)

1. C - інформація про кліки на документи.
2. WS - веб-сервер.
3. ADB - аналітична база даних, що зберігає інформацію про дії користувачів.
4. p - дані для машинного навчання, а саме оновлення векторної моделі представлення слів.
5. VSM - модель векторного представлення слів для розширення слів.
6. v - векторний простір, у якому здійснюється пошук слів.
7. F - матриця ознак.

Структура елементу інвертованого індексу

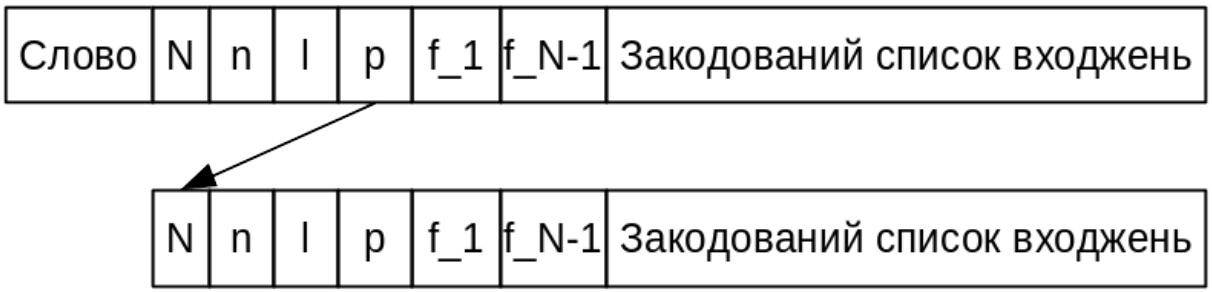

Рис.: Діаграма робити системи мультимодального пошуку

Причини вибору такої структури:

1. Файлові системи не працюють із мільярдами файлів.
2. Кількість входжень терму у колекцію є попередньо невідомою.

Структура елементу інвертованого індексу - умовні позначення

1. N - кількість входжень слова у документи.
2. n - кількість доступних байт для списку входжень.
3. l - кількість байт, зайнятим списком входжень.
4. d_1 - номер першого документи у списку вдходжень.
5. d_{n-1} - номер останнього номеру в списку входень.
6. p - індекс наступного положення у пам'яті.

Ймовірнісна функція релевантності

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \quad (1)$$

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}, \quad (2)$$

1. $k_1 = 1.5$ - величина вкладу кожного члену.
2. $b = 0.75$ - коефіцієнт нормалізації довжини документу.

Види застосування ймовірнісної моделі в залежності від поняття терму

- словах (модель мішка слів);
- символьних n-грамах;
- словесних n-грамах

Крайні значення BM25

$$\lim_{f(q_i,D)\to\infty} score_{D,q_i}(q_i, D) = k_1 + 1 \tag{3}$$

$$\lim_{f(q_i,D)\to\infty} score_{D,Q} = n(k_1 + 1) \tag{4}$$

Недоліки моделі на основі мішка слів на прикладі

Колекція документів, запит: “C better Rust”

1. Why C++ is better than C. (BM25 = $\ln\frac{4}{3} + \ln\frac{4}{3} = \ln\frac{16}{9}$)
2. Why C is better than Rust. (BM25 = $\ln\frac{4}{3} + \ln\frac{4}{3} + \ln\frac{4}{2} = \ln\frac{32}{9}$)
3. Why Rust is better than C. (BM25 = $\ln\frac{4}{3} + \ln\frac{4}{3} + \ln\frac{4}{2} = \ln\frac{32}{9}$)
4. Other random document from the collection. ($BM25 = 0$)

Недоліки гіпотези про мішок слів

1. Не захищена від описок.
2. Не захоплює семантику тексту.

Переваги

1. Може ефективно використовуватися для пошуку множини документів-кандидатів.
2. Слугує основою для тематичного моделювання.

Схема навчання з учителем

1. Дано документи фіксованої структури із пошуковими запитами та мірою релевантності.
2. Ознаки будуються для кожного поля на основі ймовірнісних функцій релевантності.
3. В якості алгоритму навчання використовується ансамбль дерев (Градієнтний бустинг).

Отримання вибірки для навчання векторного представлення слів

1. Користувач вводить запит.
2. Система надає результати користувачу для перегляду.
3. Користувач ідентифікує релевантні документи.
4. Запит розширюється словами із знайдених релевантних документів, що мають найбільшу вагу.
5. Нові терми додаються до пошукового запиту.

Швидкодія алгоритмів пошуку елементів у багатовимірному розрідженому просторі

Табл.: Кількість запитів за секунду при розмірності, рівній 256

Стурктура даних	Кількість запитів за секунду
KD-Tree	5
Hilbert R-Tree	10
LSH	103

Приклад пратичного застосування

```
<database name="store">
  <table name="question">
    <column name="title" type="uint8_t" n="128">
      <index type="inverted_term_int64_t"></index>
      <index type="inverted_ngram_int64_t" n="3"></index>
    </column>

    <column name="description" type="uint8_t"  n="128">
      <index table="question" column="description" type="inverted_term_int64_t"></index>
      <index type="inverted_ngram_int64_t" n="3"></index>
    </column>

    <column name="brand" type="uint8_t" n="128">
      <index table="question" column="brand" type="rb_tree"></index>
    </column>

    <column name="created_at" type="uint64_t"></column>
  </table>
  <index table="question" column="created_at" type="rb_tree"></index>
</database>
```

Рис.: Таблиця питань у базі Інтернет-магазину

Приклади записів із бази

Табл.: Таблиця товарів

Заголовок	**Опис**	**Бренд**
Simpson Strong-Tie 12-Gauge Angle	Not only do angles make joints stronger, they also provide	whirlpool
BEHR Premium Textured DeckOver 1-gal. SC-141 Tugboat Wood and Concrete Coating	BEHR Premium Textured DECKOVER is an innovative ...	Behr
Whirlpool 1.9 cu. ft. Over the Range Convection Microwave in Stainless Steel with Sensor Cooking	Achieving delicious results is almost ...	Whirlpool

Приклади запитів у контексті документу

Табл.: Пошукові запити користувачів

id	Текст заголовку	Текст запиту	R
1	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3
1	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.5
2	1804 Dual Spray Half Pattern 4 in Pop-Up Spray Head	rainbird sprinkler	2.33
3	Samsung 4.2 cu. ft. Front Load Washer with Steam in White ENERGY STAR	PLATFORM FOR WASHERS	2.67
3	Nantucket Pavers Patio-on-a-Pallet 10 ft. x 10 ft.	Belgium block pavers	2

Зв’язок міри Жаккара із релевантністю

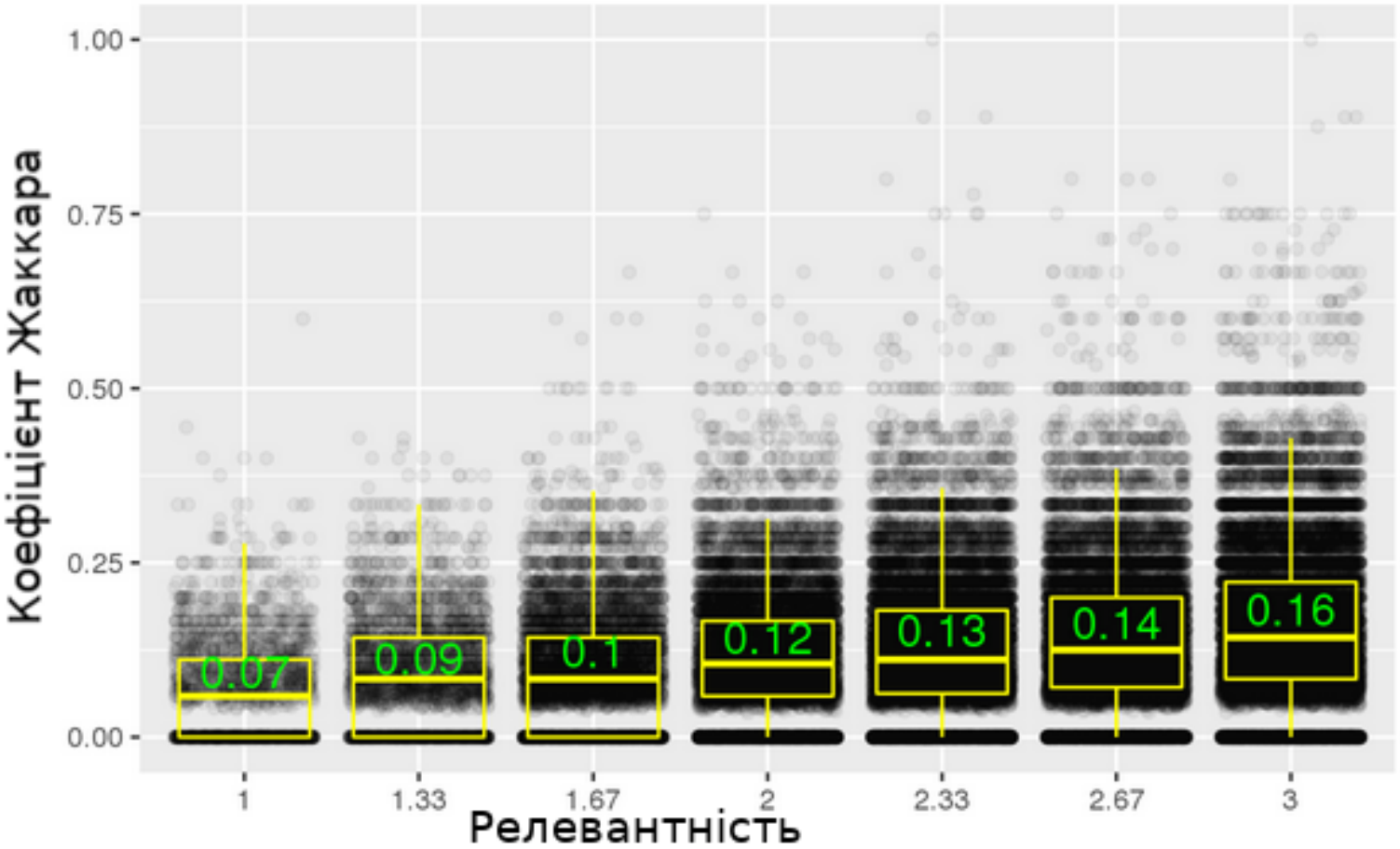

Табл.: Коірнь із середньоквадратичне відхилення

Алгоритм	$\sqrt{\sigma}$
Дерево рішень (без розширення набору ключових слів)	0.51
Дерево рішень (з розширенням набору ключових слів)	0.50
Ліс рішень (без розширення набору ключових слів)	0.459
Ліс рішень (з розширенням набору ключових слів)	0.445
CatBoost (без розширення набору ключових слів)	0.456
CatBoost (з розширенням набору ключових слів)	0.435

Висновки

Результат роботи

- Розроблено систему мультимодального пошуку.
- Розкриття інформаційної невизначеності досягається застосуванням векторної моделі слів.
- Розкриття критеріальної невизначеності - за допомгою ансамбля дерев.

Перспективи подальших

- Індвивідуальна адаптація пошукової видачі.
- Додавання нових ознак для покращення моделі машинного навчання.
- Точне налаштування гіперпараметрів моделі.