УДК 004.8

Т. О. Савчук
А. П. Коханівський

СТРУКТУРА ІНТЕЛЕКТУАЛЬНОГО МОДУЛЯ ПОШУКУ РІЗНОФОРМАТНИХ ДАНИХ В БАЗАХ ДАНИХ

Вінницький національний технічний університет

Анотація

Запропонована структура інтелектуального модуля пошуку різноформатних матеріалів у базах даних дозволяє підвищити точність знайдених результатів завдяки використанню семантики в пошуковому двигуні.

Ключові слова: семантичний пошук, система пошуку, обробка природної мови, машинне навчання.

Abstract

The proposed structure of the intelligent search module for various types of materials in databases allows for increased accuracy of search results by utilizing semantics in the search engine.

Keywords: semantic search, search system, natural language processing, machine learning.

У сучасному світі, де потоки інформації невпинно зростають, пошук релевантних даних стає дедалі складнішим завданням. Перенавантаження інформацією призводить до ситуації, коли знайти потрібну інформацію стає все важче. Саме тому виникає нагальна потреба в розробці інтелектуальних систем пошуку, які могли б ефективно опрацьовувати різноформатні дані та надавати користувачам точні й релевантні результати.

Традиційні системи пошуку, засновані на збігах ключових слів, виявляються недостатньо точними та релевантними [1]. Семантичний пошук, заснований на аналізі контексту й змісту запиту, є перспективним рішенням для покращення якості пошукових систем. Таким чином, побудова інтелектуального модуля зі структурою, що спиратиметься на семантичний пошук матеріалів, є більш точною з погляду точності отриманих результатів, оскільки він розглядатиме саме контекстуальну репрезентацію запиту та шуканої інформації.

Сучасні системи пошуку будуються на основі комбінування декількох підходів, що дозволяє створити потужну й гнучку систему пошуку. Така система здатна ефективно опрацьовувати складні запити та надавати максимально релевантні результати. Серед основних підходів є пошук за ключовими словами, алгоритми ранжування результатів, персоналізований пошук та семантичний пошук [2]. Розглянемо ці підходи детальніше:

1. Пошук за ключовими словами:

Пошук за ключовими словами - найпоширеніший метод, за якого користувач вводить слова чи фрази для знаходження релевантної інформації. Система повертає результати, що містять ці ключові слова у своєму вмісті або метаданих. Хоча цей базовий підхід може бути неточним та не враховувати контекст запиту, його простота, ефективність та гнучкість роблять пошук за ключовими словами корисним як основним, так і допоміжним інструментом у різних пошукових системах. У простих системах він може бути єдиним методом, тоді як у складніших доцільно його комбінувати з іншими підходами для підвищення релевантності результатів [2]. Цей підхід підходить, як допоміжний інструмент, оскільки він дозволяє знизити ймовірність отримання нерелевантних результатів завдяки додатковому співставленню слів, що, у свою чергу, підвищить загальну точність отриманих результатів.

2. Персоналізований пошук:

Пошукові системи можуть адаптувати результати для конкретного користувача на основі його історії, уподобань, місцезнаходження тощо. Це підвищує релевантність, проте створює ризик "фільтр-бульбашки". Персоналізований пошук набув популярності через зростання обсягів даних і потребу фільтрації. Використовуючи штучний інтелект та статистичні алгоритми, він відфільтровує масиви інформації, залишаючи релевантну для користувача. Існують різні методи імплементації.

Рекомендується поєднувати цей підхід з іншими для оптимізації результатів і мінімізації "фільтр-бульбашки" [2]. Однак, через необхідність отримувати зворотний зв'язок від користувача, цей підхід обмежений у застосуванні.

3. Семантичний пошук:

Сучасні пошукові системи застосовують технології обробки природної мови, комп'ютерного зору та семантичного аналізу для кращого розуміння запитів користувачів і надання більш релевантних результатів на основі виявлення їхнього справжнього наміру та контексту, а не лише пошуку за ключовими словами [2]. Нині семантичний пошук реалізується переважно на основі метаданих, що не завжди точно відображає семантику матеріалу. Це ускладнює отримання семантично релевантної інформації. Повноцінна імплементація потребує значних обчислювальних ресурсів, через що часто використовуються спрощені алгоритми на шкоду точності [3]. Враховуючи переваги семантичного пошуку в розумінні запитів та їх контексту, а також можливість використання потужних обчислювальних ресурсів для забезпечення високої точності, цей підхід має високий потенціал в якості основного методу пошуку.

Переваги та недоліки цих підходів наведено в таблиці 1.

Таблиця 1 – Переваги та недоліки підходів пошуку різноформатних даних в сучасних пошукових системах

Підхід	Переваги	Недоліки
Пошук за ключовими словами	- Простий і поширений; - Ефективний та гнучкий;	- Може бути неточним; - Не враховує контекст запиту;
Персоналізований пошук	- Підвищує релевантність результатів - Популярний через зростання обсягів даних	- Ризик "фільтр-бульбашки" - Потребує зворотного зв'язку від користувача
Семантичний пошук	- Краще розуміння запитів та контексту - Надає більш релевантні результати - Можливість використання потужних ресурсів для високої точності	- Повноцінна імплементація потребує значних ресурсів

Аналізуючи результати, наведені в таблиці, можна дійти висновку, що використання семантичного пошуку як основного алгоритму в поєднанні з пошуком за ключовими словами дозволить виконувати задачу з вищою точністю. При цьому, семантичний аналіз матеріалів забезпечить релевантність результатів, а удосконалений алгоритм пошуку за ключовими словами допоможе їх коригуванню та підвищенню загальної точності [3].

Удосконалений алгоритм пошуку різноформатних даних в базах даних включає кроки:

1. Обробка даних:

Даний сервіс необхідний для функціонування прогрманого засобу, оскільки без нього складність інтелектуального модуля значно зростає через потребу опрацювання набору форматів вхідних даних. Цей сервіс дозволяє не лише працювати з різними форматами, але й здійснювати їх подальше опрацювання:

- Зміна формату вхідних матеріалів до текстового, для подальшого опрацювання.
- Приведення інформації до векторного вигляду для швидкого пошуку й ефективного зберігання в базі даних.
- Отримання ключових слів з тексту й приведення їх до векторів.
- Збереження в базі даних.

Опрацювання даних у такому вигляді спростить розробку та подальше масштабування інтелектуального модуля й підвищить продуктивність програмного засобу порівняно з його відсутністю.

2. Пошук даних за семантикою:

Цей сервіс є основною частиною усього модуля, оскільки саме завдяки ньому здійснюється основне завдання – пошук. Складність даного сервісу є мінімальною, адже все, що потрібно зробити в його межах, – це перевірити подібність векторів:

- Отримання векторної репрезентації запиту та отримання доступу до бази даних.

— Повний перебір бази даних для пошуку векторів, подібних до вектора запиту.
— Запис оцінки схожості в результати.

Семантичний пошук даних у межах цього модуля є найпростішим серед усіх інших завдань, оскільки він не потребує розробки складних алгоритмів й дозволяє досить швидко проводити пошук матеріалів за рахунок використання таких алгоритмів, як косинусна подібність.

3. Пошук за ключовими словами:

У даній імплементації пошук за ключовими словами є допоміжним інструментом, що дозволятиме фільтрувати нерелевантні результати, які семантичний пошук оцінив як подібні із запитом через обмежену кількість інформації або типовість запиту:

— Отримання векторної репрезентації ключових слів із запиту та отримання доступу до бази даних.
— Повний перебір бази даних для пошуку подібних векторів.
— Запис оцінки схожості в результати.

Цей сервіс також можна використовувати як основний метод пошуку інформації, оскільки в такій імплементації ми отримуємо семантичний зміст слова, але через обмеженість потенціалу пошуку з використанням лише семантики слова, використання даного підходу буде обмеженим.

4. Відображення результатів:

Для відображення результатів буде використано веб-інтерфейс, розроблений для даного модуля. Для коректного відображення результатів буде отримано значення подібності шляхом перемноження оцінок подібності на коефіцієнти, а результати будуть відсортовані відповідно до їхньої оцінки подібності:

— Користувацький інтерфейс.
— Отримання загального значення подібності результатів із запитом.
— Сортування результатів пошуку та відображення на веб-інтерфейсі користувача.

Даний сервіс дозволить взаємодію з користувачем.

Реалізація удосконаленого алгоритму передбачає наявність таких блоків у структурі відповідного інтелектуального модулю:

1. Блок зміни формату вхідної інформації: В межах даного компонента відбувається обробка вхідних данів різних форматів, для подальшого опрацювання у вигляді текстової інформації. Також цей блок дозволяє спростити розробку та можливість розширення модуля. Цей блок взаємодіє з блоками обробки ключових слів з тексту та обробки тексту.
2. Блок обробки ключових слів з тексту: Цей блок дозволяє отримати ключові слова з тексту й привести їх у вигляд вектора для подальшого використання. Взаємодія відбувається між блоками запису інформації в базу даних та блоком пошуку даних.
3. Блок обробки тексту: Цей компонент виконує схожу функцію до минулого блоку, але замість отримання ключових слів, відбувається отримання векторної репрезентації всього тексту. В даному блоці взаємодія також відбувається між блоками запису інформації в базу даних та блоком пошуку даних.
4. Блок запису інформації в базу даних: Цей компонент дає можливість запису інформації в базу даних. Дані, що записані даним компонентом мають наступну структуру: оригінальні дані, їх векторна репрезентація, ключові слова, векторна репрезентація ключових слів. Даний блок не передає інформації та не взаємодіє з іншими блоками.
5. Блок пошуку даних: Основний елемент всього модуля. Він дозволяє здійснювати пошук по отриманій раніше з запиту та матеріалів в базі даних інформації. В процесі пошуку кожному з матеріалів проставляється число, що відповідає схожості між запитом та матеріалом. Відбувається взаємодія між даним блоком та блоком обробки результатів.
6. Блок обробки результатів: Після здійснення процесу пошуку та отримання оцінок схожості, проводиться обчислення фінального коефіцієнту схожості між результатами та запитом. Ці дані сортуються й передаються на сторону клієнта. Цей блок взаємодіє з блоком відображення результатів.
7. Блок відображення результатів: Відображення результатів здійснюється через веб-інтерфейс та дозволяє ознайомитись з результатами пошуку. Даний блок не передає інформації та не взаємодіє з іншими блоками.

Блок зміни формату вхідної інформації переводить вхідні дані до текстового формату, й передає їх на обробку ключових слів з тексту, а також на компонент обробки тексту, які, в свою чергу, отримують

відповідну інформацію та передають її на запис у випадку отримання нових матеріалів, або, на блок пошуку даних – у випадку запиту. Після проведення пошуку відбувається обробка результатів в межах відповідного компонента. Оброблені дані передаються на веб-інтерфейс, де користувач може отримати результати.

Структурна схема відповідного інтелектуального модуля, що включає означені блоки, має вигляд представлений на рисунку 1.

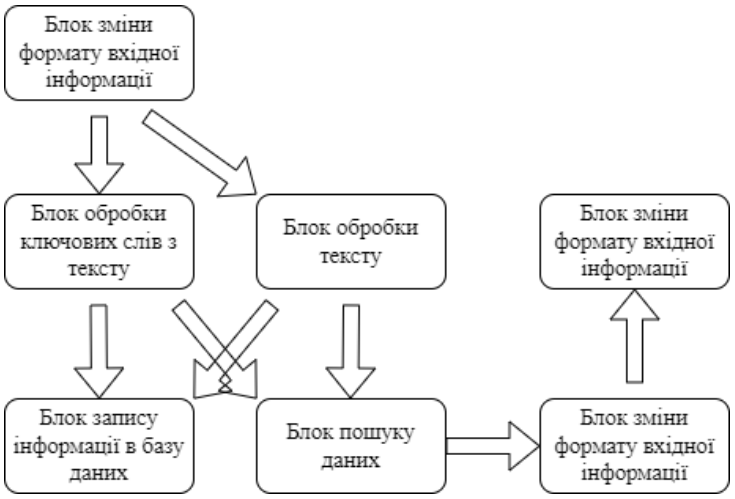

Рисунок 1 – Структурна схема взаємодії блоків інтелектуального модуля пошуку різноформатних даних в базах даних

Отже, запропонована структура інтелектуального модуля пошуку різноформатних матеріалів в базах даних забезпечить підвищення точності пошуку даних в базах даних завдяки використанню контекстуальної інформації окремих слів, текстів та різноформатних матеріалів в процесі пошуку, а також удосконаленого алгоритма пошуку за ключовими словами, що дозволяє підвищити релевантність отримуваних результатів.

Список використаної літератури

1. Ліу, Б. (2011). Семантичний пошук на основі онтологій. Видавництво Спрінгер.
2. Манінгер, Д., & Векслер, Д. (2019). Обробка природної мови для семантичного пошуку. Журнал досліджень в галузі штучного інтелекту, 35(2), 123-156.
3. Гупта, С., & Гупта, А. (2022). Машинне навчання для семантичного пошуку: сучасний стан та перспективи. Огляд інформатики та комунікацій, 18(4), 321-345.

Савчук Тамара Олександрівна — PhD, професор кафедри комп'ютерних наук Вінницький національний технічний університет, м. Вінниця, e-mail: savchtam@gmail.com

Коханівський Антон Павлович — студент групи 1КН-20б, факультет інтелектуальних інформаційних технологій та автоматизації Вінницький національний технічний університет, Вінниця, e-mail: balalauka62@gmail.com

Savchuck Tamara Olexandrivna *— PhD, Professor of the Computer Sciences Chair, Vinnytsia National Technical University, Vinnytsia, e-mail: savchtam@gmail.com*

Kokhanivskyi Anton Pavlovich *— student of group 1kn-20b, faculty of intellectual information technologies and automation, Vinnytsia National Technical University, Vinnytsia, e-mail:* balalauka62*@gmail.com*