УДК 004.048

В.В. Семенюк
Тернопільський національний технічний університет імені Івана Пулюя, Україна

ПОВНОТЕКСТОВИЙ ПОШУК КОНТЕНТУ КОНСОЛІДОВАНОГО СОЦІОКОМУНІКАЦІЙНОГО ІНФОРМАЦІЙНОГО РЕСУРСУ «РОЗУМНОГО МІСТА»

V.V. Semeniuk

FULL-TEXT SEARCH CONTENT CONSOLIDATED SOCIO-COMMUNICATION INFORMATION RESOURCE SMART CITY

Пошук є одним з ключових варіантів використання системи для консолідації соціокомунікаційних інформаційних ресурсів «Розумного міста». Реалізація відповідних функціональних наборів, поданих на діаграмах в публікації [1], з забезпеченням видачі результату найбільш релевантного до критеріїв запиту, є актуальним завданням реалізації множини функціональних можливостей системи. З метою забезпечення результативного пошуку, на першому етапі необхідно виділити основні елементи, які впливають на його якість (рис. 1).

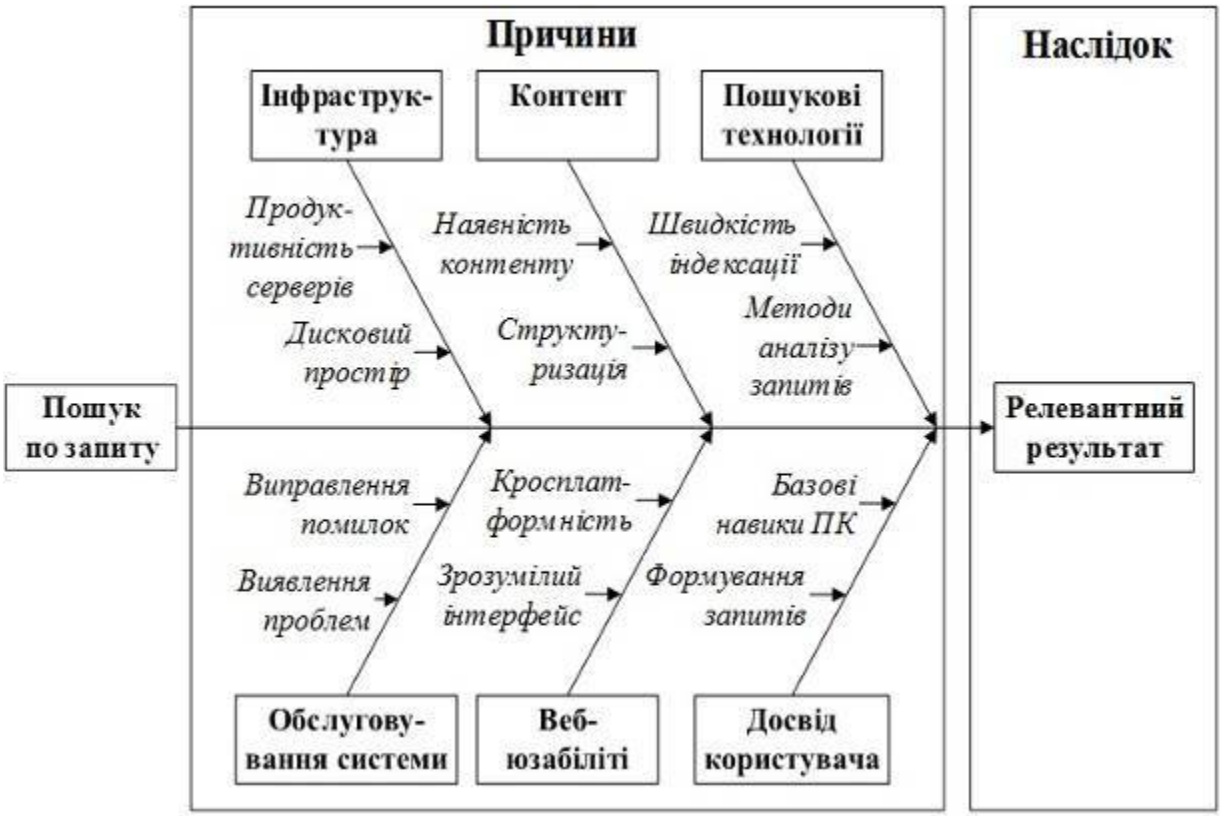

Рисунок 1. Причинно-наслідкова діаграма якості пошуку

Розглядаючи детальніше особливості реалізації пошукових алгоритмів, можна виділити два основних підходи до організації пошуку на веб-ресурсах:

1. Точний. Передбачає точний пошук по вмісту БД (в конкретних полях) без аналізу запиту.
2. Повнотекстовий. Цей підхід можна віднести до напряму Natural Language Processing, пошук здійснюється по всіх словах запиту у всіх частинах документу з використанням додаткових засобів (синонімічних баз, морфологічних словників і т.д.),

що збільшує ймовірність отримання потрібного результату навіть за умови нечіткого формулювання запиту. Крім цього, можна коригувати запит з допомогою автодоповнення чи пропозицій (при помилковому написанні слів).

Перший варіант ефективно використовується для веб-ресурсів з невеликою кількістю контенту, котрий, при потребі, можна швидко знайти за допомогою засобів навігації. У випадку глобального пошуку по одиницях контенту всіх дочірніх ресурсів консолідованого інформаційного ресурсу це значно знижує продуктивність БД та результативність.

Другий підхід дозволить генерувати максимальну вибірку з множиною результатів та повертати найбільш релевантний результат на основі аналітичного опрацювання запиту з врахуванням уточнюючих критеріїв та метрик, таких як діапазон дат, регіони, рейтингові параметри, тематики, тип елементів контенту, стоп-слова, наявність мультимедійного контенту у вмісті документу та інших.

Враховуючи типи контенту консолідованого соціокомунікаційного інформаційного ресурсу «Розумного міста», пошук повинен передбачати можливість генерації вибірки не тільки з опублікованих елементів контенту, а і з коментарів та завантажених документів. Тому для реалізації пошукових алгоритмів доцільним буде застосування другого підходу.

Впровадження одного з існуючих програмних рішень для організації повнотекстового пошуку з подальшою адаптацією до особливостей консолідованого соціокомунікаційного інформаційного ресурсу «Розумного міста» пришвидшить час розробки та відлагодження зазначеного функціоналу. Тому доцільно використати пошуковий рушій, який дозволить також зменшити навантаження на БД, збільшивши точність та швидкість пошуку. Оскільки пошуковий рушій здійснює індексацію вмісту документів та зберігає їх, виконуючи попереднє опрацювання аналізатором, то для його розгортання та безперебійної роботи потрібно забезпечити достатню кількістю обчислювальних та фізичних ресурсів.

Провівши порівняльну оцінку існуючих пошукових рушіїв Elasticsearch, Sphinx та Solr [2], вирішено використати Elasticsearch, який містить додаткові засоби візуалізації даних (за допомогою плагіна Kibana), що покращить функціональні можливості централізованого моніторингу та адміністрування [3].

В подальшому необхідно спроектувати структуру БД консолідованого соціокомунікаційного інформаційного ресурсу «Розумного міста», яка забезпечуватиме функціонування системи синхронно з пошуковими алгоритмами та програмними засобами.

Література

1. Пасічник В. В., Кунанець Н. Е., Дуда О. М., Липак Г. І., О Мацюк. В., Семенюк В. В. Актори та діаграми прецедентів системи консолідації соціокомунікаційних інформаційних ресурсів "Розумних міст". Науковий вісник НЛТУ України. 2017. Вип. 27(10). С. 129–136.

2. Max L. Elasticsearch vs. Solr vs. Sphinx: Best Open Source Search Platform Comparison - Greenice [Електронний ресурс] / Max Lapko // Greenice. – 2018. – Режим доступу до ресурсу: https://greenice.net/elasticsearch-vs-solr-vs-sphinx-best-open-source-search-platform-comparison/.

3. Сбор и анализ логов и метрик распределенного приложения с помощью Elasticsearch/Logstash/Kibana (ELK) [Електронний ресурс]. – 2014. – Режим доступу до ресурсу: http://2014.secrus.org/2014/files/040_sukhorukov.pdf.