УДК 002.537

ПОШУКОВА СИСТЕМА ДЛЯ БАЗИ ДАНИХ "УФД БІБЛІОТЕКА"

Бачеріков В. В.
Науковий керівник – доц., к.т.н. Майданюк В. П.

Задачею розробки було створення ефективної пошукової системи для бази даних “УФД Бібліотека”. Пошукова система в асимптотиці повинна мати O(n) час індексування, та O(1) для виконання пошукового запиту. Також пошук мав враховувати морфологію, тобто пошук слова “Книга”, знайде “Книги” також.

Постановка задачі пов’язана з тим, що стандартний модуль веб-інтерфейсу пошуку у базі даних, який постачається з програмним комплексом "УФД Бібліотека" має ряд недоліків. Хоча він і виконує фасетний пошук за автором, назвою, типом та мовою документа, але не досить ефективно використовує ресурси комп'ютера. Пошук у великій базі даних стає повільним і може перевантажити комп'ютер. Також він не дозволяє шукати одразу декілька типів документів. Для ефективного пошуку на даний час існує багато рішень, таких як Apache Solr, Elasticsearch, Sphinx, Whoosh та інші. Кожен з них має свої переваги і недоліки. Так Elasticsearch єє одним з кращих рішень для пошуку даних які змінюються в реальному часі. Також він легко маштабується. Ще одна перевага це документо-орієнтований індекс, це позбавляє необхідності описувати структуру документу. Але документи в базі даних оновлюється не так часто, а швидкість пошуку в Elasticsearch меньша ніж в деяких інших систем. Краще всього задачі відповідає система Sphinx яка має високу швидкодію індексування, та пошуку. Також вона підтримує пошук за морфологією. До недоліків можна віднести те що система на даний час не підтримує українську морфологію, але вона підтримує російську. Для створення пошукового індексу ця система на відміну від Elasticsearch вимагає опису схеми документа. База даних "УФД Бібліотека" підтримує велику кількість СУБД, тому при розробці пошукової системи доцільно використовувати ORM. Також слід врахувати деякі особливості СУБД Firebird, яка дозволяє доступ за один раз тільки для деякої кількості полів. Тому якщо необхідно доступ для більшої кількості, запит слід розбити на дві частини. У якості алгоритмічної мови було обрано Python, як мову яка дозволяє легке та швидке проектування. Для цієї мови доступне офіційне API для Sphinx. Але існують деякі проблеми з підтримкою СУБД Firebird. На даний час існує два драйвера для нього, але один є застарілим, а інший експериментальний, тобто ще незакінчений. Для ефективної роботи з базою та індексом результати запитів слід кешувати. Для цього доцільно використовувати таку систему як memcached. Вона дозволяє швидке кешування, та розподілення на велику кількість комп'ютерів. Так наприклад запит до бази даних який рахує кількість документів в системі можна зберегти на деякий час в кеші, так як швидко ця кількість не зміниться.