FullText Search – Đơn giản mà hữu ích!!

LỜI NÓI ĐẦU

* Là một lập trình viên mà đã từng phải thao tác với cơ sở dữ liệu, hay đơn thuần là đã từng là một trang web bán hàng, chắc hẳn các bạn đã từng nghe qua về khái niệm “full text search”
* Khái niệm này đã được định nghĩa khá cụ thể và đấy đủ trên wikipedia. Nói một cách đơn giản, “Full text search” là kĩ thuật tìm kiếm trên “Full text database”, ở đây “Full text database” là cơ sở dữ liệu chứa “toàn bộ” các kí tự(text) của một hoặc một số các tài liệu, bài báo…(document), hoặc là của websites.
* Trong bài viết này, mình sẽ giới thiệu và Full Text Search, từ khái niệm đến ứng dụng thực tiễn của kĩ thuật này.

Introduction

* Chắc hẳn các bạn đã từng dùng qua một kĩ thuật tìm kiếm rất cơ bản, đó là thông qua câu lệnh LIKE của SQL.  
  Ex: Select column\_name(s) from table\_name where column\_name like pattern;
* Sử dụng LIKE, các bạn sẽ chỉ định tìm kiếm ở column đã định trước, do đó lượng thông tin phải tìm giới hạn lại chỉ trong các column đó.
* Câu lệnh LIKE cũng tương đương với việc bạn maching pattern cho “từng” chuỗi của từng dòng(rows) của field tương ứng, do đó về độ phức tạp sẽ là tuyến tính với số dòng, và số kí tự của từng dòng, hay chính là “toàn bộ kí tự chứa trong field cần tìm kiếm”. Do đó sử dụng LIKE query sẽ gặp các vấn đề:

1. Không chính xác

+ Độ nhiễu cao

+ Từ đồng nghĩa

+ Từ cấu tạo bằng chữ đầu của cụm từ

1. Performance không tốt (Tốc độ truy vấn chậm, %keywork% không dùng index)
2. Vấn đề với tìm kiếm tiếng Việt có dấu và không dấu

* Như vậy chúng ta cần một kĩ thuật tìm kiếm khác, tốt hơn LIKE query, mềm dẻo hơn, tốt về performance hơn, đó chính là Full text search.
* Cơ bản về kĩ thuật Full text search. Về mặt cơ bản, điều làm nên sự khác biệt giữa full text search và các kĩ thuật search thông thường khác chính là “inverted index”. Vậy đầu tiên chúng ta sẽ tìm hiểu về inverted index