Lớp: IT003.O21.CTTN

SO SÁNH CÁC THUẬT TOÁN STRING MATCHING

Thời gian thực hiện: 19/03/2024 – 26/03/2024

Sinh viên thực hiện: Nguyễn Văn Hồng Thái

MSSV: 23521418

Nội dung: So sánh ba thuật toán trong string matching

-Knuth-Morris-Pratt algorithm

-Boyer-Moore algorithm

-Rabin Karp algorithm

I. Tổng quan

Bài toán tìm kiếm xâu kí tự (string searching, hay đôi khi gọi là đối sánh xâu - string matching) là một trong những bài toán cơ bản và quan trọng trong các thuật toán xử lý về xâu ký tự hay xử lý văn bản (text processing). Đây là thuật toán xử lý xâu văn bản quan trọng và có nhiều ứng dụng trong thực tế. Có rất nhiều thuật toán tìm kiếm xâu kí tự ví dụ như thuật toán Brute Force, thuật toán Knuth - Morris- Pratt, thuật toán Karp -Rabin, Boyer-Moore,…

Ứng dụng:

-Tìm kiếm chuỗi trong một văn bản: ứng dụng trong việc tìm kiếm thông tin trên Internet của các trình duyệt như Google, FireFox, Edge,…
-Kiểm tra đạo văn.
-So khớp các kí tự để đưa ra dự đoán dựa trên như dữ liệu có sẵn
-Ứng dụng trong sinh học, đi tìm các chuỗi nucleoit của một DNA.
-Ứng dụng trong xử lý ngôn ngữ tự nhiên.
-Công cụ hỗ trợ lập trình viên.
…

Theo lý thuyết độ phức tạp của 3 thuật toán lần lượt là :
Karp-Rabin: Tồi nhất: O(m*n),tốt nhất O(m+n).
Knuth – Morris – Pratt: O(m+n).
Boyer – Moore: Tồi nhất: O(m*n), tốt nhất O(n/m).

II. So sánh

1. Knuth-Morris-Pratt algorithm

a. Ưu điểm

-Đảm bảo được độ ổn định trong trường hợp tồi nhất với thời gian tiền xử lí là O(m) và thời gian tìm kiếm là O(n).
-Dễ cài đặt, và độ chính xác cao.
-Hiệu quả trong đoạn văn bản mà chuỗi con lập lại nhiều lần.

b. Nhược điểm:

-Cần thêm thời gian và không gian ở giai đoạn tiền xử lí (extra space). Điều này ảnh hưởng rất lớn đối với những văn bản có kích thước lớn.

2. Boyer-Moore algorithm

a. Ưu điểm

-Hiệu quả cho tìm kiếm trên cái đoạn văn bản có kích thước lớn không trùng lập lại.

-Đây được coi là thuật toán hiệu quả nhất nhất trong so khớp chuỗi ví dụ trong soạn thảo và thay thế văn bản.

b. Nhược điểm

-Không tốt cho các văn bản ngắn hoặc nội dung lập lại quá nhiều lần ví dụ: Binary string.

-Trường hợp tệ nhất độ phức tạp là O(m*n).

3. Rabin Karp algorithm

a. Ưu điểm

-Tìm kiếm trong chuỗi đa dạng với thời gian chạy trung bình tốt O(n+m).

-Dễ dàng thực thi

b.Nhược điểm

-Có thể có trường hợp sai do sử dụng kỹ thuật hashing.

-Chạy chậm trong tình huống xấu nhất

III. Thông tin chi tiết

https://github.com/hgthai-uit/StringMatching

(Email liên hệ: 23521418@gm.uit.edu.vn (Nguyễn Văn Hồng Thái))