ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Nguyễn Thanh Hùng

NGHIÊN CỨU PHƯƠNG PHÁP SO SÁNH XÂU XẤP XỈ VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - Năm 2015

LỜI CAM ĐOAN

Tôi xin cam đoan số liệu và kết quả nghiên cứu trong luận văn này là trung thực và chưa được sử dụng để bảo vệ học hàm, học vị nào.

Tôi xin cam đoan: Mọi sự giúp đỡ cho việc thực hiện luận văn này đã được cám ơn, các thông tin trích dẫn trong luận văn này đều đã được chỉ rõ nguồn gốc.

Thái Nguyên, ngày 18 tháng 8 năm 2015

TÁC GIẢ LUẬN VĂN

Nguyễn Thanh Hùng

LỜI CẢM ƠN

Trong thời gian nghiên cứu và thực hiện luận văn này, em đã may mắn được các thầy cô chỉ bảo, dìu dắt và được gia đình, bạn bè quan tâm, động viên. Em xin bày tỏ lời cảm ơn sâu sắc nhất tới tất cả các tập thể, cá nhân đã tạo điều kiện giúp đỡ em trong suốt quá trình thực hiện nghiên cứu luận văn này.

Trước hết em xin trân trọng cảm ơn Ban giám hiệu trường Đại học Công nghệ thông tin và truyền thông, Phòng Đào tạo và Khoa Sau đại học của nhà trường cùng các thầy cô giáo, những người đã trang bị kiến thức cho em trong suốt quá trình học tập.

Với lòng biết ơn chân thành và sâu sắc nhất, em xin trân trọng cảm ơn thầy giáo – PGS.TS Nguyễn Trí Thành, giảng viên khoa Công nghệ thông tin – Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội; người thầy đã trực tiếp chỉ bảo, hướng dẫn khoa học và giúp đỡ em trong suốt quá trình nghiên cứu, hoàn thành luận văn này.

Xin chân thành cảm ơn tất cả các bạn bè, đồng nghiệp đã động viên, giúp đỡ nhiệt tình và đóng góp nhiều ý kiến quý báu để em hoàn thành luận văn này.

Do thời gian nghiên cứu có hạn, luận văn của em chắc hẳn không thể tránh khỏi những sơ suất, thiếu sót, em rất mong nhận được sự đóng góp của các thầy cô giáo cùng toàn thể bạn đọc.

Xin trân trọng cảm ơn!

Thái Nguyên, ngày 18 tháng 8 năm 2015

TÁC GIẢ LUẬN VĂN

Nguyễn Thanh Hùng

MỤC LỤC

DANH MỤC HÌNH VẼ

MỞ ĐẦU

1. Lý do chọn đề tài

Kiểu dữ liệu văn bản (Text) là dạng trình bày thông tin gần gũi nhất với con người, vì vậy, đây cũng là dạng trình bày thông tin số rất phổ biến. Chính vì lẽ đó, bài toán tìm kiếm văn bản (text searching) là một trong những bài toán quan trọng nhất trong hoạt động tìm kiếm thông tin của con người. Trong thời đại ngày nay, văn bản số hóa đang tăng trưởng "bùng nổ" trong các cơ sở dữ liệu trên Internet, dung lượng tăng gấp đôi sau mỗi chu kỳ 18 tháng. Trong bối cảnh đó, vấn đề tìm kiếm văn bản một cách tự động đã rất quan trọng thì lại ngày càng quan trọng hơn.

Dạng phổ biến nhất của bài toán tìm kiếm văn bản là: Cho trước nguồn tìm kiếm là một tập ***D*** các văn bản (hoặc là cơ sở dữ liệu văn bản, hoặc là tập các văn bản trên Internet). Cho một câu hỏi dạng văn bản ***q*** (thường là một từ, một xâu văn bản ngắn), hãy tìm tất cả các văn bản thuộc ***D*** mà có chứa ***q***. Trong nhiều trường hợp (chẳng hạn, tìm kiếm thông qua máy tìm kiếm) thì ***q*** còn được gọi là "truy vấn" và bài toán còn có tên gọi là "tìm kiếm theo truy vấn". Để tìm được các văn bản có chứa văn bản truy vấn ***q***, hệ thống tìm kiếm cần phải kiểm tra văn bản truy vấn ***q*** có là một xâu con của các văn bản thuộc tập ***D*** hay không (sánh mẫu) và đưa ra các văn đáp đáp ứng. Trong nhiều trường hợp, bài toán còn đòi hỏi tìm tất cả các vị trí của các xâu con trong văn bản trùng với ***q***. Đồng thời, điều kiện tìm kiếm có thể được làm "xấp xỉ" theo nghĩa văn bản kết quả có thể không cần chứa ***q*** (không cần có một xâu con của văn bản trùng một cách hoàn toàn chính xác với ***q***) mà chỉ cần "liên quan" tới ***q*** (có xâu con trong văn bản "xấp xỉ" ***q***). Có thể thấy, các máy tìm kiếm sử dụng cả cơ chế tìm kiếm xấp xỉ khi mà văn bản kết quả tìm kiếm không chứa hoàn toàn chính xác văn bản truy vấn [1].

Thời gian gần đây, bài toán sánh mẫu càng trở nên quan trọng và được quan tâm nhiều do sự tăng trưởng nhanh chóng của các hệ thống tìm kiếm thông tin và các hệ thống sinh- tin học. Một lý do nữa, con người ngày nay không chỉ đối mặt

với một lượng thông tin khổng lồ mà còn đòi hỏi những yêu cầu tìm kiếm ngày càng phức tạp. Các mẫu đưa vào không chỉ đơn thuần là một xâu ký tự mà còn có thể chứa các ký tự thay thế, các khoảng trống và các biểu thức chính quy. Sự "tìm thấy" không đơn giản là xuất hiện chính xác mẫu trong văn bản mà còn cho phép "một xấp xỉ" giữa mẫu và xuất hiện của nó trong văn bản. Từ đó, bên cạnh vấn đề kinh điển là "tìm kiếm chính xác", nảy sinh một hướng nghiên cứu là *"sánh mẫu xấp xỉ / tìm kiếm xấp xỉ"* (approximate matching / approximate searching) [2].

So sánh thực nghiệm của thời gian chạy của thuật toán xấp xỉ chuỗi kết hợp cho *k* vấn đề khác nhau được trình bày. Với một chuỗi mô hình, một chuỗi văn bản, và một số nguyên *k*, nhiệm vụ là để tìm tất cả các lần xuất hiện gần đúng của mô hình trong văn bản với ít nhất *k* khác biệt (chèn thêm, xóa, thay đổi). Xem xét bảy thuật toán dựa trên phương pháp tiếp cận khác nhau bao gồm lập trình năng động, Boyer-Moore chuỗi kết hợp, hậu tố bị tự động, và sự phân bố của các nhân vật. Nó chỉ ra rằng không ai trong số các thuật toán là tốt nhất cho tất cả các giá trị của các thông số vấn đề, và sự khác biệt tốc độ giữa các phương pháp có thể là đáng kể.

Xuất phát từ những yêu cầu và lý do trên, em lựa chọn đề tài luận văn là: **"Nghiên cứu phương pháp so sánh xâu xấp xỉ và ứng dụng"**.

Luận văn này định hướng nghiên cứu một số thuật toán so sánh mẫu xâu xấp xỉ, tập trung vào một số thuật toán của Galil Park, Ukkonen Wood Boyer-Moore, thuật toán xâu vòng tròn gần đúng với độ phức tạp là hàm tuyến tính và tiến hành thực nghiệm ứng dụng.

2. Mục tiêu nghiên cứu

- Nghiên cứu để hiểu các khái niệm và đặc trưng liên quan tới bài toán so sánh xâu xấp xỉ.
- Nghiên cứu các lớp thuật toán so sánh xâu xấp xỉ.
- Khảo sát, phân tích một số thuật toán và các bước tiến hóa và hiệu suất (nghiên cứu khả năng về ý tưởng cải tiến thuật toán).

- Khảo sát chương trình ứng dụng asmf-master để có thể khai thác vào trường hợp của luận văn.

- Cài đặt thử nghiệm tìm vị trí các câu hỏi trong nội dung, kết quả trả về của một máy tìm kiếm.

3. Đối tượng và phạm vi nghiên cứu

Nghiên cứu một số thuật toán so sánh xâu xấp xỉ miền dữ liệu văn bản. Tiếp đó, luận văn thi hành một số thuật toán trong họ thuật toán nói trên, cài đặt thử nghiệm tìm kiếm.

4. Ý nghĩa thực tiễn của luận văn

Nghiên cứu thuật toán so sánh xâu xấp xỉ và ứng dụng của chúng vào hệ thống tìm kiếm văn bản. Vì vậy, nó có ý nghĩa rất lớn trong lý thuyết và thực tiễn.

5. Phương pháp nghiên cứu

- Phương pháp nghiên cứu tài liệu, phân tích, tổng hợp.

- Phương pháp thực nghiệm và đối chứng qua chương trình thử nghiệm.

CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ XẤP XỈ

Trong chương này sẽ trình bày về một số số khái niệm xấp xỉ, khái quát về một số thuật toán đối sánh mẫu xấp xỉ, các giải pháp thực hiện cho áp dụng ứng dụng cho thuật toán đối sánh chuỗi xấp xỉ mà tác giả đã nghiên cứu trong thời gian vừa qua.

1.1 Khái niệm xấp xỉ

Khái niệm xấp xỉ trong luận văn này, được sử dụng đồng nghĩa với khái niệm đối sánh chuỗi xấp xỉ. Luận văn này sẽ trình bày một số giải thuật về đối sánh chuỗi xấp xỉ mà tác giả đã nghiên cứu và tìm hiểu được các ứng dụng thực tế như tìm và sửa từ lỗi, gợi ý tìm kiếm trong hệ thống tìm kiếm từ điển. Đối sánh thể hiện việc so sánh chuỗi ***T*** và chuỗi ***P***. Các kỹ thuật đối sánh được ứng dụng nhiều trong các lĩnh vực khác nhau của tin học. Trong luận văn này, các kỹ thuật được sử dụng để phát hiện từ sai chính tả và sử lại các từ sai.

1.1.1 Đối sánh chuỗi

Đối sánh thể hiện việc so sánh chuỗi ***T*** và chuỗi ***P***. Các ký tự của ***T*** được so sánh với các ký tự của ***P***. Với phương pháp so sánh khác nhau sẽ trả lời các yếu tố tương quan của **T** và **P** theo góc độ của thuật toán cụ thể. Ví như đối sánh theo các thao tác chuyển ký tự của ***T*** để ***T*** chuyển thành ***P*** và ngược lại, phương pháp này thể hiện độ đo khoảng các đối sánh. Ngoài ra còn nhiều phương pháp đối sánh khác nhau cho phép đối sánh **T** và **P** theo *n_gram*, đối sánh mẫu theo tiền tố, hậu tố …

Bài toán đối sánh chuỗi là kỹ thuật thực hiện tìm kiếm sự xuất hiện của chuỗi **T** trong chuỗi **P**. Cho chuỗi **T** và chuỗi **P**, sử dụng giải thuật **F** tìm kiếm và trả về kết quả **R** là các vị trí của chuỗi **T** xuất hiện trong chuỗi **P**.

$$\boldsymbol{R = F(T,P)}$$

Như vậy có thể thấy kết quả trả về của giải thuật ***F*** phụ thuộc vào:

- Cung cấp dữ liệu đầu vào ***T***: Nếu cung cấp dữ liệu ***T*** không tốt sẽ cho kết quả không như mong đợi, thậm trí không có kết quả.

- Cung cấp dữ liệu xử lý ***P***: Chuỗi cần xử lý để đưa ra kết quả xem có sự xuất hiện của ***T*** trong ***P*** hay không. Phụ thuộc vào nguồn cung cấp ***P***. Nếu nguồn cung cấp ***P*** không chứa chuỗi ***T*** thì việc đối sánh trả về kết quả là không tìm thấy kết quả mà nguồn ***T*** cung cấp.

- Bản thân giải thuật ***F***: Giải thuật sẽ thực hiện phương pháp đối sánh ***T*** với ***P*** và trả về kết quả theo yêu cầu. Giải thuật ảnh hưởng trực tiếp đến kết quả. Tùy thuộc vào cách tiếp cận mà có kết quả khác nhau. Nếu giải thuật tiếp cận theo hướng đối sánh chính xác cần phải thực hiện sánh đúng chuỗi ***T*** chuỗi ***P*** để trả lời câu hỏi ***P*** có trùng ***T*** hay không.

Các kỹ thuật đối sánh có thể phân thành hai loại đó là kỹ thuật so sánh đối sánh chính xác, kỹ thuật còn lại là so sánh đối sánh không chính xác (xấp xỉ). Trong luận văn này chỉ đề cập tới các thuật toán xấp xỉ.

1.1.2 Đối sánh chính xác

Tìm chính xác là phương pháp trả lời câu hỏi chuỗi ***T*** có chính xác tồn tại trong chuỗi ***P*** hay không? Nếu có hãy chỉ ra vị trí xuất hiện của ***T*** có trong ***P*** [3]. Khái niệm trùng được chỉ ra ở đây có thể hiểu theo ý nghĩa là ***T*** được sánh đúng với chuỗi con của ***P*** theo từng thứ tự ký tự, độ dài của ***T*** và chuỗi con phải bằng nhau.

Ví dụ : Cho chuỗi ***T***="*Các*"

và chuỗi ***P*** ="*Chào các bạn. Chúc các bạn ngày mới tốt lành*".

Như vậy kết quả của đối sánh trả về thứ tự 1 và 4 là vị trí xuất hiện của ***T*** trong ***P***.

Đối sánh chính xác đưa ra một số ý nghĩa của nó thực hiện trong các hệ thống tìm kiếm cần trả về kết quả chính xác. Hệ thống sẽ trả lời câu hỏi ***T*** có thật sự tồn tại trong ***P*** hay không? Với các phương pháp của tìm kiếm chính xác đủ mạnh để có một hiệu quả tìm kiếm tốt.