ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

TRƯƠNG KIM TÚ

TÌM HIỂU PHƯƠNG PHÁP XỬ LÝ TÌM KIẾM THEO KÝ TỰ ĐẠI DIỆN CỦA LUCENE

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

Hà Nội – 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

TRƯƠNG KIM TÚ

TÌM HIỂU PHƯƠNG PHÁP XỬ LÝ TÌM KIẾM THEO KÝ TỰ ĐẠI DIỆN CỦA LUCENE

Ngành: Hệ thống Thông tin

Chuyên ngành: Hệ thống Thông tin

Mã số: 60.48.0104

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS Nguyễn Trí Thành

Hà Nội – Năm 2016

LỜI CẢM ƠN

Tôi muốn bày tỏ lòng biết ơn sâu sắc tới những người đã giúp đỡ tôi trong quá trình làm luận văn, đặc biệt tôi xin cám ơn PGS. TS. Nguyễn Trí Thành, với lòng kiên trì, thầy đã chỉ bảo tôi chi tiết và cho tôi những lời nhận xét quí báu trong từng bước làm luận văn. Đồng thời tôi cũng xin gửi lời cảm ơn tới các thầy cô giáo khoa Công nghệ thông tin - Trường Đại học Công nghệ - Đại học Quốc gia Hà nội đã truyền đạt các kiến thức cho tôi trong suốt thời gian học tập và nghiên cứu vừa qua.

Tôi cũng xin chân thành cảm ơn cơ quan, bạn bè, đồng nghiệp, gia đình và những người thân đã cùng chia sẻ, giúp đỡ, động viên, tạo mọi điều kiện thuận lợi để tôi hoàn thành nhiệm vụ học tập và cuốn luận văn này.

Hà Nội, tháng 5 năm 2016

Học viên

Trương Kim Tú

LỜI CAM ĐOAN

Tôi xin cam đoan nội dung trình bày trong luận văn này là do tôi tự nghiên cứu tìm hiểu dựa trên các tài liệu và tôi trình bày theo ý hiểu của bản thân dưới sự hướng dẫn trực tiếp của PGS. TS Nguyễn Trí Thành. Các nội dung nghiên cứu, tìm hiểu và kết quả thực nghiệm là hoàn toàn trung thực.

Luận văn này của tôi chưa từng được công bố trong bất cứ công trình nào. Trong quá trình thực hiện luận văn tôi đã tham khảo tài liệu của một số tác giả, tất cả những thống tin liên quan đến tài liệu tham khảo đều được liệt kê trong mục "TÀI LIỆU THAM KHẢO" ở cuối luận văn.

Tôi xin chịu trách nhiệm hoàn toàn về lời cam đoan của mình, nếu có gì sai, tôi sẽ chịu mọi hình thức kỷ luật theo quy định.

Hà Nội, tháng 5 năm 2016

Học viên

Trương Kim Tú

MỤC LỤC

MỞ ĐẦU

1. Đặt vấn đề

Ngày nay, với sự ra đời của mạng Internet và sự phát triển nhanh chóng, vượt bậc của mạng truyền thông, một khối lượng rất lớn các thông tin được cập nhật và đưa lên mạng thường xuyên. Các thông tin là các tập tin có cấu trúc hoặc phi cấu trúc, nằm rải rác ở nhiều nơi. Câu hỏi đặt ra làm thế nào để tìm được đúng thông tin một cách nhanh chóng và hiệu quả nhất. Để đáp ứng yêu cầu đó, đã có rất nhiều phương pháp tìm kiếm thông tin cũng như các công cụ tìm kiếm thông tin ra đời như Google, Yahoo, Altavista, Bing…

Tuy nhiên, thông tin cần tìm kiếm là rất nhiều và đa dạng và nhu cầu tìm kiếm thông tin của người dùng ngày càng cao nên việc nghiên cứu, tìm hiểu để khám phá và hiểu biết sâu hơn về cách thu thập, lưu trữ, biểu diễn, tổ chức tìm kiếm thông tin hiệu quả và nhanh nhất vẫn thực sự rất cần thiết.

Dựa trên nhu cầu trên rất nhiều kỹ thuật tìm kiếm cơ bản và nâng cao đã được đưa ra giới thiệu và được áp dụng trong rất nhiều công cụ tìm kiếm phổ biến hiện nay. Tuy nhiên, phạm vi nghiên cứu của luận văn chỉ dừng lại ở việc giới thiệu những nét cơ bản nhất của các phương pháp tìm kiếm phổ biến hiện nay, sau đó tập trung vào việc tìm hiểu phương pháp tìm kiếm theo ký tự đại diện từ khái quát, giải thuật cho đến cài đặt thực tế với một thư viện tìm kiếm mạnh mẽ là Lucene.

2. Mục tiêu nghiên cứu

Nghiên cứu của luận văn hướng tới các mục tiêu sau:

- Tìm hiểu về tìm kiếm nói chung và tìm kiếm theo ký tự đại diện nói riêng.
- Tìm hiểu các giải pháp tìm kiếm theo ký tự đại diện
- Tìm hiểu giải pháp tìm kiếm theo ký tự đại diện của Lucene
- Tiến hành thực nghiệm tìm kiếm theo ký tự đại diện của Lucene cho tiếng Việt

3. Cấu trúc luận văn

Luận văn được chia thành 4 phần với các nội dung như sau:

Chương 1 trình bày tổng quan về các phương pháp tìm kiếm. Các kiến thức được trình bày bao gồm các phương pháp chung được sử dụng trong tìm kiếm, đặc biệt là phương pháp tìm kiếm theo ký tự đại diện và ý nghĩa của nó về mặt khoa học và thực tiễn nhằm mang lại những kiến thức căn bản nhất trong lĩnh vực tìm kiếm.

Chương 2 Trình bày sâu hơn về kỹ thuật xử lý truy vấn và các giải thuật tìm kiếm theo ký tự đại diện. Các kỹ thuật được trình bày trong chương này sẽ là cơ sở lý thuyết cho việc tìm hiểu và cài đặt chương trình ứng dụng ở chương tiếp theo.

Chương 3 giới thiệu thư viện Lucene và tính năng tìm kiếm theo ký tự đại diện của Lucene, từ đó vận dụng vào việc xây dựng chương trình thử nghiệm tính năng tìm kiếm theo ký tự đại diện của Lucene.

Phần kết luận tổng kết những kết quả đạt được của luận văn và hướng nghiên cứu tiếp theo.

.

Chương 1. TỔNG QUAN

Chương đầu tiên của luận văn cung cấp cái nhìn tổng quan về tìm kiếm thông tin trên Internet và những thách thức hiện nay đối với vấn đề này. Để giải quyết những tốt những vấn đề gặp phải trong tìm kiếm thông tin rất nhiều phương pháp tìm kiếm từ cơ bản đến nâng cao được đề xuất, trong đó có phương pháp tìm kiếm theo ký tự đại diện. Các khái niệm cơ bản nhất của các phương pháp này sẽ được trình bày một cách ngắn gọn nhất trong nội dung chương 1.

1.1 Tổng quan về các phương pháp tìm kiếm

Internet có thể được xem như là một kho thông tin khổng lồ và vô tận, được cung cấp từ hàng triệu Website trên khắp thế giới, chứa đựng hầu như toàn bộ kiến thức của nhân loại. Trên Internet người dùng có thể tìm được vô số thông tin bổ ích và các kiến thức về mọi lĩnh vực từ khoa học cho đến lịch sử, văn học… Tuy nhiên, nguồn tri thức đó lại không được sắp xếp theo một trật tự.

Vì vậy, trước một kho thông tin như thế nếu người dùng chưa có mục đích tìm kiếm rõ ràng thì sẽ mất thời gian vì lượng thông tin quá nhiều. Thêm nữa nếu không thành thạo, người dùng sẽ rất khó khăn trong việc tìm thấy thông tin cần thiết trong lượng lớn kết quả tìm kiếm.

Chính vì thế phương pháp tìm kiếm thông tin trên Internet được xem là một kỹ năng vô cùng quan trọng và cần thiết.

Các phương pháp tìm kiếm cơ bản cần phải kể đến ở đây gồm: Tìm kiếm chính xác, tìm kiếm theo ký tự đại diện, tìm kiếm theo mệnh đề, tìm kiếm xấp xỉ và tìm kiếm cụm từ. Trong tìm kiếm chính xác, chỉ những tài liệu chứa chính xác từ khóa người dùng nhập vào được hiển thị. Còn trong trường hợp người dùng không nhớ được chính xác từ khóa tìm kiếm, tìm kiếm theo ký tự đại diện là một những giải pháp phù hợp được sử dụng vì nó sử dụng các ký tự như “?” hoặc “*” để đại diện cho không hoặc một ký tự bất kỳ hay một chuỗi ký tự bất kỳ (gồm cả chuỗi có độ dài bằng 0). Tìm kiếm theo mệnh đề có sử dụng các toán tử logic như AND, OR để liên kết các câu truy vấn đơn tạo thành một mệnh đề tìm kiếm phức tạp hơn. Để tăng tính liên quan của các tài liệu được trả về thì tìm kiếm cụm từ là một kỹ thuật hữu ích. Tìm kiếm xấp xỉ cũng là một kỹ thuật tìm kiếm hay được sử dụng trong tìm kiếm thông tin, phương pháp này sẽ trả về kết quả chứa thuật ngữ gần giống với thuật ngữ truy vấn đưa ra bởi người sử dụng.

Ngoài các phương pháp tìm kiếm cơ bản được trình bày ở trên, một số phương pháp tìm kiếm nâng cao cũng được các công cụ tìm kiếm sử dụng nhằm

làm mịn hơn kết quả tìm kiếm: Tìm kiếm tập hợp, tìm kiếm theo trường xác định.... Trong tìm kiếm theo tập hợp, kết quả tìm kiếm được hiển thị như các tập hợp, và có thể kết hợp với các tìm kiếm khác hay các từ khóa khác. Tìm kiếm theo trường cụ thể cho phép người dụng lựa chọn một trường cụ thể để thực hiện tìm kiếm thay vì thực hiện tìm kiếm với tất cả các trường.

1.2 Tổng quan về phương pháp xử lý tìm kiếm theo ký tự đại diện

Truy vấn theo ký tự đại diện được sử dụng trong những tình huống sau đây: (1) người dùng không chắc chắn về cách viết của một thuật ngữ truy vấn (ví dụ, *Sydney* với *Sidney*, sẽ dẫn đến truy vấn theo ký tự đại diện *S*dney*); (2) người dùng biết có nhiều biến thể trong cách viết của một thuật ngữ (ví dụ, *color* với *colour*); (3) người dùng tìm kiếm các tài liệu chứa các biến thể của một thuật ngữ có thể nhận được thông qua giải thuật stemming, nhưng không chắc chắn các công cụ tìm kiếm có thực hiện giải thuật stemming hay không (ví dụ, *judicial*, với *judiciary*, sẽ dẫn đến truy vấn theo ký tự đại diện *judicia**); (4) người dùng không chắc chắn về cách viết đúng của một từ hay cụm từ nước ngoài (ví dụ, truy vấn *Universit* Stuttgart*).

Các cơ sở dữ liệu, công cụ tìm kiếm khác nhau sẽ sử dụng các ký tự khác nhau làm ký tự đại diện. Tuy nhiên, dấu * và dấu ? là các ký tự đại diện được sử dụng phổ biến nhất. Trong phạm vi nghiên cứu của luận văn hai ký tự đại diện phổ biến là dấu * và dấu ? sẽ được tìm hiểu.

- Dấu * đại diện cho chuỗi ký tự bất kỳ, gồm chuỗi có độ dài bằng 0. Ví dụ:
  - *s*food* tìm kiếm: *seafood* hoặc *soyfood*
  - *enzym** tìm kiếm: *enzyme* hoặc *enzymes* hoặc *enzymatic* hoặc *enzymic*
  - *Hof*man** tìm kiếm *Hofman* hoặc *Hofmann* hoặc *Hoffman* hoặc *Hoffmann*
- Dấu ? đại diện cho không hoặc một ký tự bất kỳ. Ví dụ:
  wom?n tìm kiếm: woman hoặc women.
- Trong một truy vấn tìm kiếm có thể sử dụng kết hợp các ký tự đại diện khác nhau. Ví dụ:
  *organi?ation** tìm kiếm: *organisation* hoặc *organisations* hoặc *organisational* hoặc *organization* hoặc *organizations* hoặc *organizational*

Các cơ sở dữ liệu, công cụ tìm kiếm khác nhau sẽ có những quy tắc khác nhau trong việc tìm kiếm theo ký tự đại diện, sao cho việc thực hiện tìm kiếm đạt hiệu quả nhất. Tuy nhiên, để có thể tận dụng tối đã những lợi ích mà kỹ thuật

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Nguyễn Văn Định (2012). "*Giáo trình Otomat và Ngôn ngữ hình thức"*. NXB Đại học Nông Nghiệp.

Tiếng Anh

2. Christian Charras, Thierry Lecroq (2004), *Handbook of Exact String - Matching Algorithms,* College Publications.
3. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze (2009), *An Introduction to Information Retrieval*, Cambridge University Press, England, Online edition (c) 2009 Cambridge UP.
4. G.Berry, R.Sethi (1986), "From regular expressions to deterministic automata", *Theoretical Computer Science*, Elsevier Science Publishers B.V. (North-Holland), pp.117-126.
5. Michael McCandless, Erik Hatcher, Otis Gospodnetic (2009), *Lucene in action 2nd Edition,* Manning Publications.
6. Keneilwe Zuva, Tranos Zuva (2012), "Evaluation of Information Retrieval", *International Journal of Computer Science & Information Technology (IJCSIT)*, Vol. 4 (No. 3), June 2012.
7. Lingpipe, and Gate, Manu Konchady (2008), *Building Search Applications: Lucene*, Mustru Publishing, 1st edition.
8. Mehryar Mohri (1997), "Finite-State Transducers in Laguage and Speech Processing", *Computational Linguistics*, Volume 23 Issue 2, June 1997, pp.269-311.
9. Paul Clough, Mark sanderson (2013), "Evaluating the performance of information retrieval systems using test collections", *IR Information Research*, Vol. 18 (No. 2), June, 2013.
10. Ricardo Baeza -Yates, Berthier Ribeiro - Neto (1999), *Morden Information Retrieval,* Addison Wesley.
11. Stoyan Mihov and Denis Maurel (2001), *Direct Construction of Minimal Acyclic Subsequential Transducers*.
12. William B.Frakes, Ricardo Baeza-Yates (1992), *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, 1st edition.