Phương pháp biểu diễn ngữ nghĩa lân cận siêu liên kết trong máy tìm kiếm VietSeek

Đặng Tiểu Hùng

Người hướng dẫn: TS. Hà Quang Thụy

MỤC LỤC

LỜI CẢM ƠN

Tôi xin bày tỏ lòng kính trọng và biết ơn tới các thầy giáo, cô giáo khoa Công nghệ tr- ờng Đại học Quốc gia Hà Nội đã dìu dắt tôi trong suốt quá trình học tập và nghiên cứu, cũng nh- đóng góp những ý kiến quý báu cho luận văn.

Đặc biệt tôi xin bày tỏ lòng kính trọng và biết ơn sâu sắc Thầy giáo Tiến sĩ Hà Quang Thuỵ cùng gia đình đã tận tình, dành nhiều thời gian h- ớng dẫn, động viên, khích lệ cho tôi hoàn thành luận văn này.

Tôi xin bày tỏ lòng biết ơn sâu sắc tới gia đình, bạn bè và đồng nghiệp đã tạo điều kiện thuận lợi giúp đỡ cũng nh- có nhiều ý kiến đóng góp bổ ích cho luận văn.

Tôi xin kính chúc các thầy giáo, cô giáo cùng gia đình mạnh khoẻ, hạnh phúc; Tiếp tục sự nghiệp đào tạo cho các thế hệ học sinh, sinh viên đạt đ- ợc nhiều thành công hơn nữa trên con đ- ờng học tập và nghiên cứu khoa học.

Tôi xin chúc các bạn bè, đồng nghiệp mạnh khoẻ, thành công; áp dụng hiệu quả và sáng tạo các kiến thức đ- ợc học vào thực tiễn.

Xin trân trọng cảm ơn.

Hà Nội ngày 25/03/2004

Học viên

Đặng Tiểu Hùng

PHẦN MỞ ĐẦU

Cùng với sự phát triển mạnh mẽ của Iinternet là một sốkhối l- ợng khổng lồ dữ liệu đ- ợc phát sinh, tuy nhiên (theo thông tin từcủa tập đoàn Oracle) thì khoảng 90% dữ liệu ở dạng phi cấu trúc hoặc nửa cấu trúc. Trong khi nNhu cầu khai thác, tìm kiếm thông tin một cách chính xác trên internet đã ngày càng trở nên bức thiết hơn, do đó xuất hiện các hệ tìm kiếm theo từ khoá (cụm từ khoá) nh- Yahoo, Google ... Tuy nhiên việc tìm kiếm theo từ khoá vẫn ch- a đủ để giúp ng- ời sử dụng nhanh chóng tìm đ- ợc trang Web cần thiết vì số l- ợng kết quả trả lại rất lớn và nhiều khi chỉ là các trang Web ít có liên quan. Vì vậy các hệ thống tìm kiếm ngàycần đ- ợc cải tiến để ngày càng thông minh hơn. Xuất hiện những hệ h- ớng tới mục tiêu cụ thể nh- tra cứu thông tin về các chủ đề y tế, giáo dục, luật pháp, âm nhạc ... Tuy vậy, việc nghiên cứu các giải pháp để tìm đ- ợc mộtcác trang thông tin theo một nội dung nào đó sát với yêu cầu ng- ời sử dụng thì vẫn còn nhiều hạn chế. Đã có nhiều mô hình tìm kiếm đ- ợc đề xuất, song những mô hình lý t- ởng về mặt lý thuyết thì lại ch- a có tính khả thi khi cài đặt. Do đó, trong các hệ tìm kiếm, ng- ời ta tìm cách cải tiến các ph- ơng pháp đơn giảncó sẵn để có áp dụng trong thực tế. Luận văn này h- ớng tới việc nghiên cứu, phân tích, đánh giá kết quả của một số thuật toán tìm kiếm theo nội dung, từ đó đề xuất một ph- ơng án cải tiến để nâng cao hiệu quả về tính chính xác của nội dung cũng nh- về tốc độ.

Từ việc tìm hiểu, đánh giá và phân tích - u, nh- ợc điểm của các ph- ơng pháp tiếp cận khác nhau, dựa theo mục tiêutrên ý t- ởng nâng cao hiệu quả tìm kiếm, luận văn đề xuất giải pháp thực hiện "***Ph□ơng pháp biểu diễn ngữ nghĩa lân cận siêu liên kết cho máy tìm kiếm VietSeek***".

Nội dung của luận văn đ- ợc định h- ớng vào các vấn đề sau:

1. Mô hình toán học củabiểu diễn trang văn bản Web.
2. Khái quát các ph- ơng pháp tiếp cận trong tìm kiếm trang Web có nội dung t- ơng tự. Đánh giá - u điểm và -nh- ợc điểm của mỗi ph- ơng pháp đ- ợc khảo sát.

3. Đề xuất ph- ơng pháp kết hợp để ~~đạt~~nâng cao hiệu quả ~~cao hơn~~ trong tìm kiếm trang Web có nội dung t- ơng tự.

Luận văn bao gồm Phần mở đầu, ba ch- ơng nội dung và Phần kết luận với~~mà~~ nội dung các ch- ơng đ- ợc trình bày nh- d- ới đây.

Ch- ơng 1 với tiêu đề là ***Tổng quan về các ph□ơng pháp biểu diễn và tìm kiếm thông tin trên web*** giới thiệu khái quát về các ph- ơng pháp biểu diễn và tìm kiếm trên web.

Tiêu đề của ch- ơng 2 là ***Ph□ơng pháp biểu diễn trang web theo ngữ nghĩa lân cận siêu liên kết***. Ch- ơng này ~~sẽ~~ trình bày cơ sở, nội dung của ph- ơng pháp đ- ợc đề xuất ~~cũng nh- trình~~và đánh giá ph- ơng pháp đ- ợc đề xuất với các ph- ơng pháp khác. Luận văn cũng trình bày chi tiết~~các đánh~~ các lựa chọn đ- ợc đề xuất trong mỗi b- ớc của ph- ơng pháp, từ đó chọn ra giải pháp tốt nhất.

Ch- ơng 3 ***Máy tìm kiếm VietSeek và thử nghiệm Thuật toán tìm kiếm theo ngữ nghĩa lân cận siêu liên kết*** giới thiệu kiến trúc logic của máy tìm kiếm VietSeek, thiết kế logic về dữ liệu theo biểu diễn vector và thuật toán tìm kiếm theo nội dung trên cơ sở biểu diễn trang web do luận văn đề xuất. ~~Trong c~~Ch- ơng này cũng đề xuất những cải tiến khi áp dụng vào thực tế để nâng cao hiệu suất thực hiện của ph- ơng pháp biểu diễn.

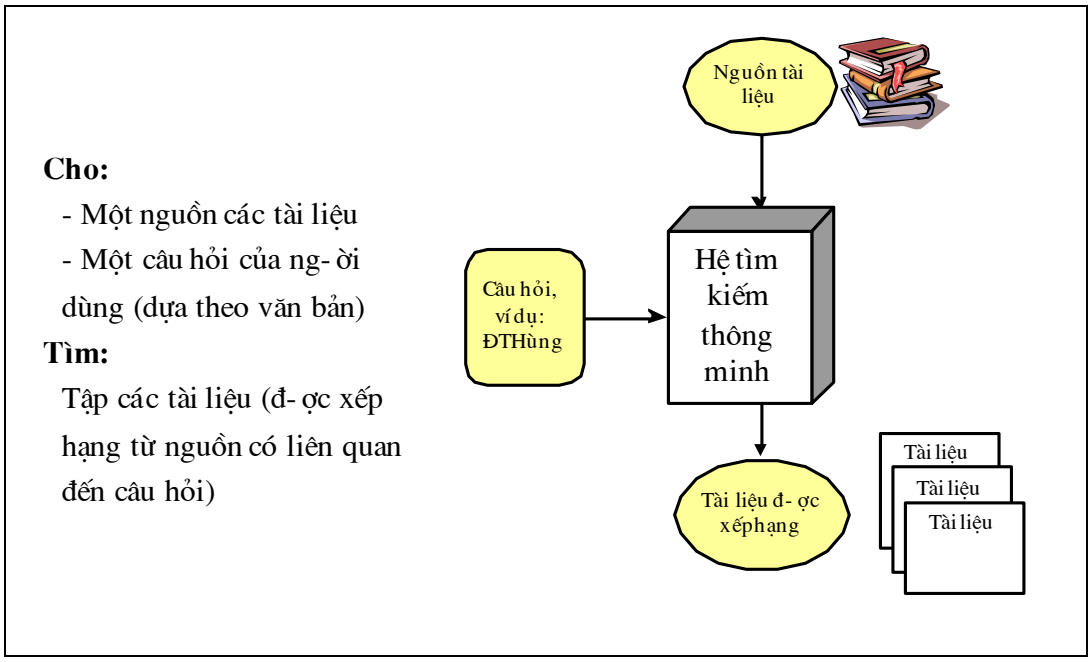
Phần kết luận tổng hợp những kết quả nghiên cứu chính của luận văn~~,~~ và chỉ ra một số hạn chế của luận văn. Đồng thời luận văn ~~cũng~~ đề xuất một số h- ớng nghiên cứu cụ thể tiếp theo của luận văn.

Phần phụ lục bổ sung một số thông tin ~~về~~ chi tiết về việc áp dụng thuật toán cho máy tìm kiếm VietSeek nh- sơ đồ khối một số module cần bổ sung chức năng, những lệnh bổ sung vào cơ sở dữ liệu của VietSeek.

CH□ONG 1. TỔNG QUAN VỀ TÌM KIẾM THÔNG TIN TRÊN WEB

1.1 Giới thiệu về tìm kiếm thông tin

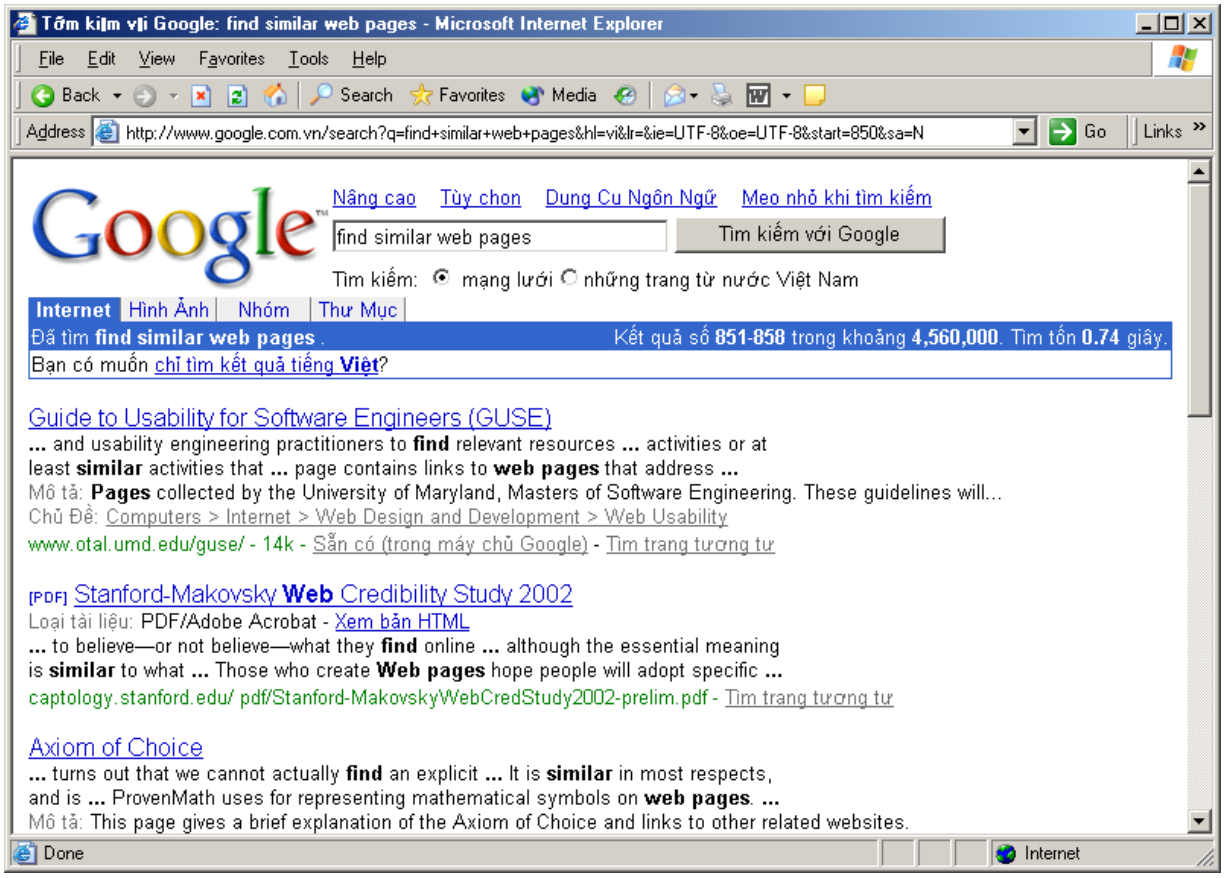
Khai phá dữ liệu~~thông tin~~ trên web (web mining) là quá trình khảo sát và phân tích dữ liệu web một cách tự động hoặc bán tự động để phát hiện ra thông tin. Từ thông tin đ- ợc khai phá, ~~và~~ tìm kiếm thông tin (Infomartion Retrieval) trên web là ph- ơng pháp để truy cập một cách hiệu quả nhất đến thông tin mà ng- ời dùng quan tâm, ~~đó có thể là~~kỳ vọng cung cấp một tập hợp nhỏ các văn bản gần nhất đến lĩnh vực hoặc chủ đề mà ng- ời dùng mong muốn tiếp cận.



Hình 1.÷Tìm kiếm thông tin

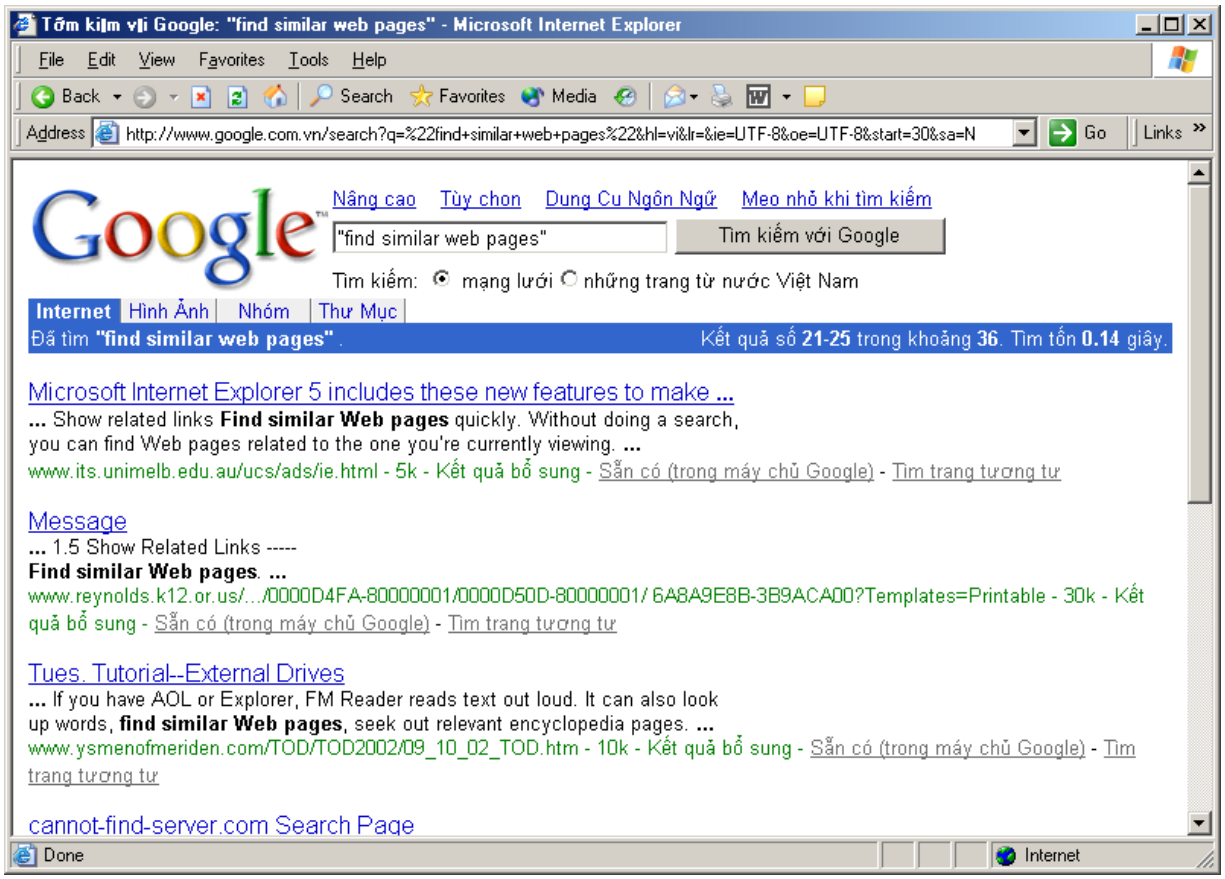
1.2 Bài toán tìm kiếm thông tin

Có 2 bài toán cơ bản trong tìm kiếm thông tin là tìm kiếm theo từ khoá và tìm kiếm theo nội dung. Bài toán tìm kiếm theo từ khoá là bài toán tìm kiếm thông tin theo các từ khóa do ng- ời dùng cung cấp [1]. Hệ tìm kiếm sẽ trả về cho ng- ời dùng các trang web có chứa những từ khoá trong câu hỏi. Tuy vậy, với số l- ợng khổng lồ các trang web trên internet nh- hiện nay thì số l- ợng kết quả tìm đ- ợc theo từ khoá là quá lớn. Ví dụ nếu tìm các trang web có từ khoá *find similar web page* thì cho kết quả 858 trang web.



Hình 2.÷ Tìm kiếm thông tin theo từ khoá

Bằng cách tìm kiếm theo cụm từ khoá thì số l- ợng kết quả trả về chính xác hơn, số kết quả trả về là 25 trang web.



Hình 3.÷ Tìm kiếm thông tin theo cụm từ khoá

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. Phạm Thanh Nam (2003). *Một số giải pháp cho bài toán tìm kiếm trong cơ sở dữ liệu Hypertext*. Luận văn thạc sĩ Công nghệ thông tin - Đại học Quốc gia Hà Nội.

[2]. Phạm Thanh Nam, Bùi Quang Minh, Hà Quang Thuỵ (2004). *Giải pháp tìm kiếm trang Web t- ơng tự trong máy tìm kiếm VietSeek*. Tạp chí Tin học và Điều khiển học (nhận đăng 1-2004).

[3]. Đoàn Sơn (2002). *Các ph- ơng pháp biểu diễn và ứng dụng trong khai phá dữ liệu văn bản*. Luận văn thạc sĩ Công nghệ thông tin - Đại học Quốc gia Hà Nội.

Tiếng Anh

[4]. J. Dean and M. Henzinger (1999). *Finding Related Pages in the World Wide Web*. Proceedings of WWW8, 1999.

[5]. L. A. Goodman and W. H. Kruskal (1954). *Measures of association for cross classifications*. J. of Amer. Stat. Assoc, 1954.

[6]. T.H. Haveliwala, A. Gionis, and P. Indyk (2000). *Scalable Techniques for Clustering the Web*.Informal Proceedings of the International Workshop on the Web and Databases, WebDB*, 2000.*

[7]. J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke (2000). *WebBase: A Repository of Web Pages*.Proceedings of WWW9, 2000.

[8]. A.K. Jain, M. Narasimha Murty, and P.J. Flynn (1999). *Data clustering: A review ACM Computing Surveys*, 31(3), 1999.

[9]. H. P. Luhn. *The Automatic Creation of Literature Abstracts*. IBM Journal of Research and Development, 2:159-165, 1958.

[10]. Nguyen Ngoc Minh, Nguyen Tri Thanh, Ha Quang Thuy, Luong Song Van, Nguyen Thi Van (2001). *A Knowledge Discovery Model in Full-text*

Databases. Proceedings of the First Workshop of International Joint Research: "Parallel Computing, Data Mining and Optical Networks". March 7, 2001, Japan Advanced Institute of Science and Technology (JAIST), Tatsunokuchi, Japan, 59-68.

[11]. M. Porter (1980). *An Algorithm for Suffix Stripping*. Program: Automated Library and Information Systems, 14(3):130-137, 1980.

[12]. G. Salton and M.J. McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[13]. Sen Slattery (2002). *Hypertext Classification*. Doctoral dissertation (CMU-CS-02-142). School of Computer Science. Carnegie Mellon University.

[14]. S. Siegel and N. J. Castellan (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988.

[15]. M. Steinbach, G. Karypis, and V. Kumar (2000). *A comparison of document clustering techniques*. TextMining Workshop, KDD, 2000.

[16]. Taher H. Haveliwala, Aristides Gionis, Dan Klein, Piotr Indyk (2002). *Evaluating Strategies for Similarity Search on the Web*. WWW2002 - USA.

[17]. BBC. http://www.bbc.com.

[18]. CNN http://www.cnn.com.

[19]. Open Directory Project (ODP). http://www.dmoz.com.

[20]. Web page www.*InfoWorld.com* (Theo công bố ngày 17/02/2004 thì trong kho dữ liệu của Google đã có 4,28 tỷ trang web, 880 triệu hình ảnh và 845 triệu thông điệp Internet. Mảng thông tin đang tăng nhanh gần đây là các trang web liên quan đến sách, bao gồm các ch- ơng đầu, phần phê bình, tham khảo. Hệ thống thông tin này đ- ợc Google truy xuất qua dịch vụ Google Print đang đ- ợc vận hành thử nghiệm. Số liệu thống kê gần đây của Google là 3,3 tỷ trang web đ- ợc kết nối vào tháng 8-2003, là 400 triệu hình ảnh vào tháng 11/2002).

[21]. Yahoo! http://www.yahoo.com/.