Kỹ thuật tìm kiếm văn bản trên cơ sở nội dung trong cơ sở dữ liệu đa phương tiện

Nguyễn Thị Thu Trang

Trường Đại học Công nghệ
Luận văn Thạc sĩ ngành: Hệ thống Thông tin; Mã số: 60 48 05
Người hướng dẫn: PGS.TS Đặng Văn Đức
Năm bảo vệ: 2010

Abstract: Giới thiệu tổng quan về cơ sở dữ liệu đa phương tiện, xếp hạng tài liệu và các yếu tố cơ bản phục vụ cho việc tìm kiếm thông tin. Khái quát về một hệ thống truy tìm thông tin (IR) tiêu biểu và cụ thể là truy tìm tài liệu văn bản. Đề cập đến vấn đề chỉ mục tài liệu và thước đo hiệu năng. Nghiên cứu một số mô hình tìm kiếm như: Boolean, không gian vectơ, phân cụm, dựa trên xác suất, phản hồi phù hợp và LSI. Tổng quan về cơ sở dữ liệu đa phương tiện, hệ thống tìm kiếm đa phương tiện đến kỹ thuật chỉ mục, xử lý tài liệu, trích lọc thông tin đến chi tiết vấn đề tìm kiếm trên tài liệu văn bản. Đặc biệt, nghiên cứu các mô hình tìm kiếm và đi sâu nghiên cứu mô hình LSI- tìm kiếm văn bản trên cơ sở nội dung.

Keywords: Công nghệ thông tin; Cơ sở dữ liệu; Hệ thống thông tin; Văn bản; Đa phương tiện

Content

MỞ ĐẦU

Hàng nghìn năm trước con người đã nhận thức được tầm quan trọng của việc lưu trữ và tìm kiếm thông tin. Với sự phát triển của máy tính, việc máy tính có khả năng lưu trữ thông tin với số lượng lớn và tìm kiếm thông tin có ích từ các tập hợp trở nên cần thiết. Lĩnh vực truy tìm thông tin (Information Retrieval - IR) ra đời vào những năm 1950 vì nhu cầu thiết yếu này. Hơn 40 năm sau, lĩnh vực đó trưởng thành đáng kể, nhiều hệ thống IR được sử dụng phổ biến với sự đa dạng trạng thái của người sử dụng. Sự phát triển của lĩnh vực này trong những năm 1970 đến những năm 1980 dựa trên nền tảng của những năm trước đó, nhiều mô hình thực hiện truy tìm tài liệu khác nhau được phát triển và tiến bộ theo mọi khía cạnh của quá trình truy tìm. Những mô hình kỹ thuật mới được chứng minh qua thực nghiệm, có hiệu quả trong những tập hợp văn bản nhỏ, có thể dùng cho các nhà nghiên cứu ở thời gian đó. Tuy nhiên, vì không có hiệu quả đối với những tập hợp văn bản lớn, câu hỏi có hay không những mô hình và những kỹ thuật có thể đáp ứng được với thể lớn hơn vẫn chưa được trả lời. Sự thay đổi lớn vào năm 1992, với sự khởi đầu bằng cuộc thảo luận về truy tìm văn bản, sau đó một loạt thảo luận kiểm định đứng đầu bởi nhiều hãng khác nhau của Mỹ dưới sự bảo hộ của Viện Tiêu chuẩn và Công nghệ quốc gia (NIST), nhằm vào việc khuyến khích nghiên cứu

về hệ thống IR với những tập hợp văn bản lớn. Những thuật toán IR đã phát triển trong những năm từ năm 1996 đến năm 1998, là những kỹ thuật đầu tiên được dùng cho việc tìm kiếm trên mạng toàn cầu.

Ngày nay, sự phát triển nhanh chóng của lĩnh vực thông tin và Internet đã tạo ra một khối lượng thông tin vô cùng lớn với sự phong phú, đa dạng và phức tạp của loại hình thông tin như: văn bản, hình ảnh, video, siêu văn bản, đa phương tiện… Tương ứng với khối lượng dữ liệu khổng lồ đó, người ta quan tâm nhiều đến cơ sở dữ liệu đa phương tiện (Mutimedia Database) trong khoa học công nghệ và trong thực tiễn. Với hệ thống cơ sở dữ liệu đa phương tiện, bao gồm dữ liệu dạng hình ảnh, video, audio và văn bản (text) đang có xu thế thâm nhập vào rất nhiều lĩnh vực và đang dần trở thành hệ cơ sở dữ liệu được quan tâm từ người sử dụng và các chuyên gia trong vấn đề lưu trữ, xử lý và ứng dụng.

Cho đến nay, vấn đề tìm kiếm thông tin đa phương tiện vẫn được các chuyên gia nghiên cứu, trong việc truy tìm thông tin phù hợp với yêu cầu của một truy vấn đưa ra từ người sử dụng. Người sử dụng có xu hướng tìm kiếm chủ yếu trong hệ cơ sở dữ liệu đa phương tiện, ví dụ như tìm kiếm một loạt hình ảnh cổ vật liên quan đến nền văn hoá cổ Việt Nam, tìm kiếm dữ liệu âm thanh có bản text kèm theo, tìm kiếm video bài giảng cho học sinh ôn thi đại học... Để thực hiện được việc tìm kiếm đó trong cơ sở dữ liệu đa phương tiện thì những người làm khoa học đã nghiên cứu ra các công cụ, phương pháp, kỹ thuật tìm kiếm sao cho thuận tiện, chính xác và nhanh chóng đem lại được thông tin phù hợp với yêu cầu của người sử dụng.

Văn bản là một trong số các dạng của dữ liệu đa phương tiện, nó được quan tâm từ hàng nghìn năm trước trong việc tổ chức sắp xếp và lưu trữ, điển hình như bảng nội dung của một cuốn sách. Ngày nay, sự lớn mạnh của thông tin với phần lớn là dạng văn bản, hơn nữa nó xuất phát từ nhu cầu thực tế sử dụng của con người. Tài liệu văn bản chiếm đa số trong mọi cơ quan tổ chức, đặc biệt là trong thư viện và còn được sử dụng để mô tả các dạng khác của dữ liệu đa phương tiện như video, audio, hình ảnh. Số lượng tài liệu văn bản ngày càng lớn và có vai trò vô cùng quan trọng, vì thế việc việc lưu trữ, xử lý và truy tìm thủ công trước đây không thể hoặc khó có thể thực hiện được. Cùng với sự ra đời và phát triển của máy tính, các công cụ xử lý cũng ngày càng hoàn thiện dựa trên những kỹ thuật hiện đại phục vụ cho nhu cầu đó.

Các mô hình truy tìm hay được sử dụng trong phạm vi này, đó là: Đối sánh chính xác, không gian vectơ, xác suất và trên cơ sở cụm. Song, nhược điểm cơ bản của các mô hình truy tìm thông tin hiện nay là những từ mà người tìm kiếm sử dụng, thường không giống với những từ đã được đánh chỉ mục trong thông tin tìm kiếm. Vấn đề này liên quan nhiều đến hai khía cạnh thực tế, đó là tính đồng nghĩa (synonymy)- cùng một thông tin nhưng được miêu tả bằng các từ khác nhau, phụ thuộc vào ngữ cảnh hay mức độ cần thiết, ví dụ như: *nhìn, xem, trông, thấy* có cùng ý nghĩa; và tính đa nghĩa (polysemy) – cùng một từ có nhiều ý nghĩa khác nhau trong ngữ cảnh khác nhau, ví dụ như: *đi* (có thể là chỉ chuyển động hay chỉ sự mất mát). Kết quả truy tìm có thể gồm những tài liệu không liên quan, đơn giản vì những thuật ngữ xuất

hiện ngẫu nhiên trong nó giống với thuật ngữ trong truy vấn và mặt khác, những tài liệu liên quan có thể bị bỏ qua bởi không chứa các thuật ngữ xuất hiện trong truy vấn (do tính đồng nghĩa). Một ý tưởng thú vị xem liệu việc truy tìm có thể dựa vào các khái niệm có hiệu quả hơn so với truy tìm trực tiếp trên các thuật ngữ. Mô hình LSI (Latent Semantic Indexing) ra đời, là một giải pháp hữu hiệu cho vấn đề truy tìm thông tin dựa trên cơ sở nội dung tài liệu văn bản, tìm kiếm trên cơ sở những khái niệm (không phải trên các thuật ngữ đơn).

Trước khi truy tìm, các tài liệu được coi như danh sách các từ và chúng phải được đánh chỉ mục. Có một thực tế là không phải tất cả các từ đều có ý nghĩa, vì vậy việc loại đi danh sách các từ không có nghĩa vô cùng quan trọng và các từ không có ý nghĩa sẽ không được đánh chỉ mục. Từ thông tin tóm lược của người sử dụng biểu thị qua truy vấn, thuật toán truy tìm phải đảm bảo rằng, chiến lược xếp hạng tập các tài liệu trong câu trả lời luôn ưu tiên cho những thông tin có ích và phù hợp với truy vấn người sử dụng đưa ra. Hơn thế nữa, một kỹ thuật được đánh giá là tốt phải dựa trên việc xếp hạng các tài liệu này, tức là những tài liệu phù hợp và được coi là "gần" với câu truy vấn nhất sẽ được xếp lên trên các tài liệu ít phù hợp hơn trong danh sách tài liệu trả lời. Đánh giá chất lượng IR còn phụ thuộc vào thước đo hiệu năng thực hiện của kỹ thuật đó dựa vào các tham số chủ yếu là độ chính xác (precison) và số tài liệu được gọi lại (recall).

Trên cơ sở đó, cấu trúc luận văn gồm phần mở đầu, kết luận, tài liệu tham khảo và phần nội dung gồm ba chương và được trình bày theo thứ tự sau:

Chương 1. Giới thiệu tổng quan về cơ sở dữ liệu đa phương tiện, xếp hạng tài liệu và các yếu tố cơ bản phục vụ cho việc tìm kiếm thông tin. Khái quát về một hệ thống truy tìm thông tin (IR) tiêu biểu và cụ thể là truy tìm tài liệu văn bản.

Chương 2. Đề cập đến vấn đề chỉ mục tài liệu và thước đo hiệu năng. Nghiên cứu một số mô hình tìm kiếm như: Boolean, không gian vectơ, phân cụm, dựa trên xác suất, phản hồi phù hợp và LSI.

Chương 3. Cài đặt thực nghiệm mô hình LSI.

Nội dung luận văn đi từ tổng quan về cơ sở dữ liệu đa phương tiện, hệ thống tìm kiếm đa phương tiện đến kỹ thuật chỉ mục, xử lý tài liệu, trích lọc thông tin đến chi tiết vấn đề tìm kiếm trên tài liệu văn bản. Đặc biệt, nghiên cứu các mô hình tìm kiếm và đi sâu nghiên cứu mô hình LSI- tìm kiếm văn bản trên cơ sở nội dung.

References

1. **Tiếng Việt**
2. PGS.TS. Đặng Văn Đức (2004-2008), *Bài giảng Cơ sở dữ liệu đa phương tiện*.
3. **Tiếng Anh**
4. Karl Aberer (2003), *Data Mining*, Laboratoire de systèmeses d'informations répartis.
5. Ricardo Baeza, Berthier Ribeiro (1999), *Modern Information Retrieval*, ACM Press New York.
6. Jamie Callan (2008), *Information Retrieval*, Carnegie Mellon University.
7. Soumen Chakrabarti (2003), *Mining the Web*, Morgan Kaufmann Publishers.

8. Scott Deerwester et al (1990), *Indexing by Latent Semantic Analysis*, Journal of The American Society for Information Science.
9. Edel Garcia (2006), *Latent Semantic Indexing (LSI) A Fast Track Tutorial*, Grossman and Frieder's Information Retrieval, Algorithms and Heuristics.
10. David Hand, Heikki Mannila & Padhraic Smyth (2001), *Principles of Data Mining*, The MIT Press, pp. 267-287.
11. Chris Manning et al (2007), *Information Retrieval and Lantent Semantic Indexing*, Lecture Notes, Marcus Uneson.
12. E.G.M Petrakis, *Multimedia Information Retrieval*, University of Maryland.
13. Gerard Salton, Chris Buckley (1988), *Parallel text search methods*, Communications of the ACM.
14. Marcel Worring, *Multimedia Information Systems*, Lecture Notes, University of Amsterdam.
15. Justin Zobel, Alistair Moffat (2006), *Inverted File for Text Search Engines*, ACM Computing Surveys, Volume. 38.
16. **Các trang web tham khảo:**
17. http://www.miislita.com/information-retrieval-tutorial/svd-lsi-tutorial-3-full-svd.html
18. http://www.bluebit.gr/matrix-calculator/