Tổng quan về truy tìm thông tin

Quách Đình Hoàng

10/07/2011

1 Giới thiệu

Ngày nay, do sự bùng nổ về thông tin nên việc tìm kiếm thông tin một cách chính xác và nhanh chóng ngày càng trở thành một nhu cầu quan trọng. Một lĩnh vực của khoa học máy tính liên quan nhiều nhất đến việc nghiên cứu và phát triển các kỹ thuật tìm kiếm là *truy tìm thông tin* (*information retrieval*). *Truy tìm thông tin* là lĩnh vực quan tâm đến việc *biểu diễn*, *lưu trữ*, *tổ chức* và *truy xuất* vào thông tin [Baeza-Yates and Ribeiro-Neto, 1999], nói một cách dễ hiểu, đó là một khoa học nền tảng cho các *công cụ tìm kiếm* (*search engine*) [Zhai, 2008]. Bài viết này giới thiệu ngắn gọn về truy tìm thông tin, đặc biệt là thông tin ở dạng *văn bản* (*text*).

Truy tìm thông tin là một lĩnh vực này nghiên cứu và phát triển những lý thuyết, nguyên lý, thuật toán và những hệ thống giúp người dùng tìm được thông tin (thường dưới dạng tài liệu văn bản) thỏa mãn nhu cầu của họ (thường được diễn đạt dưới dạng một câu truy vấn) từ một nguồn thông tin (thường rất lớn) được lưu trữ trên máy tính [Manning et al., 2008]. Tuy nhiên, theo nghĩa rộng, lĩnh vực này nghiên cứu các vấn đề giúp con người quản lý và khai thác thông tin nói chung như *tìm kiếm văn bản* (*text retrieval*), *phân loại văn bản* (*text classification*), *gom cụm văn bản* (*text clustering*), *tóm tắt văn bản* (*text summarization*), *trả lời câu hỏi*

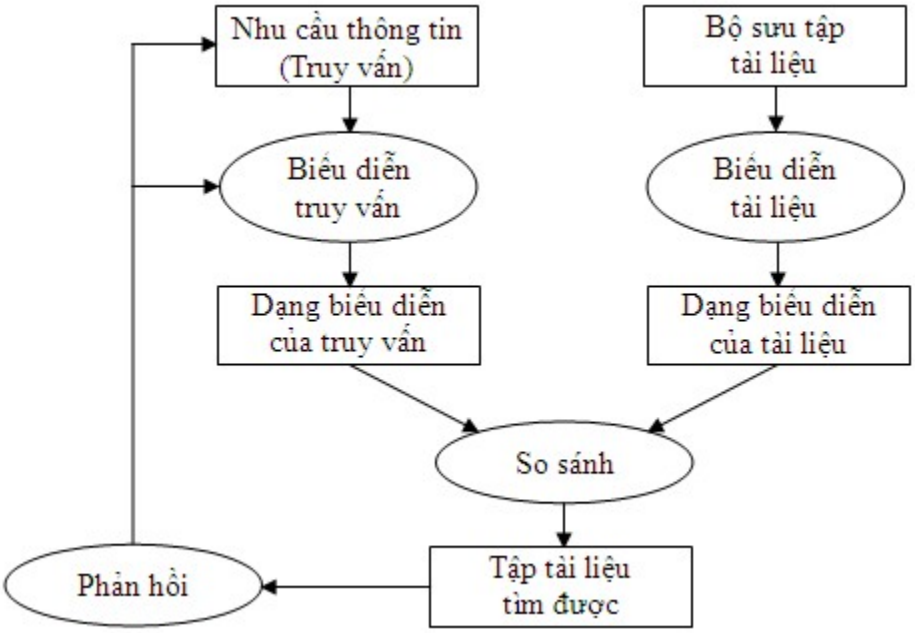
(*question answering*), *tìm kiếm văn bản đa ngôn ngữ* (*cross-language text retrieval*), *tìm kiếm thông tin đa phương tiện* (*multimedia retrieval*) như *hình ảnh*, *âm thanh*, *video*, $\cdots$ [Zhai, 2008].

2 Hệ thống truy tìm thông tin

Một hệ thống truy tìm thông tin được xây dựng để tìm kiếm trên một bộ sưu tập tập tài liệu (document collection) nhất định, đó có thể là một tập các tài liệu trên máy tính cá nhân (*desktop search engine*), một thư viện số (*digital library*), hoặc toàn bộ dữ liệu trên World Wide Web (*web search engine*).

Hình 1 minh họa 3 quá trình chính của một hệ thống truy tìm thông tin, gồm: biểu diễn truy vấn, biểu diễn nội dung tài liệu và phương pháp so sánh những biểu diễn này để quyết định tài liệu nào là phù hợp với truy vấn. Quá trình *biểu diễn truy vấn* thường được gọi là *thành lập truy vấn* (*query formulation*). Quá trình *biểu diễn tài liệu* thường được gọi là *lập chỉ mục tài liệu* (*document indexing*). Thông thường, các quá trình này chủ yếu liên quan đến việc trích xuất các từ khóa quan trọng, đại diện cho nội dung của truy vấn, tài liệu và xác định *trọng số* (*weight*) cho chúng (mức độ quan trọng của các từ khóa đó trong truy vấn, tài liệu) một cách phù hợp. Kết quả của các quá trình này là các biểu diễn bên trong của hệ thống cho mỗi truy vấn và tài liệu. Quá trình *so sánh* giữa biểu diễn tài liệu và biểu diễn truy vấn để quyết định tài liệu nào phù hợp với truy vấn được gọi là *chiến lược tìm kiếm* (*retrieval strategy*). Một chiến lược tìm kiếm sẽ xác định một giá trị thực phản ánh mức độ phù hợp giữa tài liệu với truy vấn cho mỗi tài liệu.

Tuy nhiên, truy tìm thông tin thường không phải là quá trình một chiều: truy vấn được xác định, tài liệu được trả về, người dùng có thể tìm được tài liệu mong muốn hoặc xác định một truy vấn khác. Thực tế, việc biểu diễn nhu cầu thông tin của người dùng là một quá trình động, liên quan đến những đánh giá, phản hồi của



Hình 1: Hệ thống truy tìm thông tin

người dùng về các tài liệu tìm được (xem hình 1). Nhu cầu thông tin thật sự của người dùng cũng có thể thay đổi dựa trên những đánh giá này. Dựa vào phản hồi của người dùng về những tài liệu thực sự liên quan đến truy vấn (*relevance feedback*), hệ thống tự động tạo một truy vấn mới, hy vọng tốt hơn, và quá trình tìm kiếm được lặp lại bằng cách sử dụng truy vấn mới. Tuy nhiên, trong thực tế, người dùng thường không muốn cung cấp các phản hồi. Vì vậy, một *phản hồi giả* (*pseudo relevance feedback*) có thể được thực hiện bằng cách giả định rằng những tài liệu nằm ở đầu danh sách tài liệu tìm được là có liên quan đến truy vấn. Cả hai quá *trình thành lập truy vấn* và *phản hồi* là một phần của một quá trình tương tác liên tục giữa người dùng và hệ thống để đạt được nhu cầu thông tin thật sự của người dùng.

3 Mô hình truy tìm thông tin

Mục đích chính của bất kỳ một hệ thống tìm kiếm thông tin là xếp hạng các tài liệu theo thứ tự giảm dần *mức độ liên quan* (*relevance*) với *nhu cầu thông tin của người dùng* (*user information needs*), thường dưới dạng một câu truy vấn, và loại bỏ những tài liệu *không liên quan* (*non-relevent*). Để đạt được mục đích này, hệ thống phải đưa ra một cách đánh giá mức độ liên quan giữa một *tài liệu* (*document*) bất kỳ với *câu truy vấn* (*query*) thông qua một *chiến lược tìm kiếm* (*retrieval strategy*) hay *hàm tìm kiếm* (*relevant/retrieval/ranking function*) nào đó. Chất lượng của hệ thống (khả năng cho kết quả chính xác) phụ thuộc trực tiếp vào chất lượng của chiến lược tìm kiếm này. Việc tìm một chiến lược tìm kiếm tốt luôn là một thách thức lớn trong việc xây dựng một hệ thống tìm kiếm thông tin.

Các *chiến lược tìm kiếm* được định nghĩa bởi các *mô hình truy tìm thông tin* (*information retrieval model*). Mỗi *mô hình truy tìm thông tin* là một *biểu diễn hình thức* (*formal representation*) xác định cách biểu diễn các tài liệu, các truy vấn, và phương pháp đánh giá mức độ liên quan giữa tài liệu với truy vấn [Baeza-Yates and Ribeiro-Neto, 1999]. Do đó, *độ chính xác* (*effectiveness*) của một hệ thống tìm kiếm thông tin phụ thuộc vào tính hợp lý của mô hình truy tìm thông tin bên dưới nó. Các mô hình truy tìm thông tin được nghiên cứu nhiều là *mô hình không gian vectơ* (*vector space model*) [Salton et al., 1975], *mô hình xác xuất* (*probabilistic model*) [Robertson and Jones, 1976], *mô hình mạng suy diễn* (*inference network model*) [Turtle and Croft, 1989], *mô hình ngôn ngữ* (*language model*) [Ponte and Croft, 1998] và gần đây là mô hình *learning to rank* [Burges et al., 2005, Cao et al., 2006]. Các mô hình này khi được tối ưu đều có kết quả tìm kiếm tốt gần như nhau [Zhai, 2008]. Tuy nhiên, do có cơ sở toán học vững chắc, *mô hình ngôn ngữ* và *mô hình learning to rank* được quan tâm nhiều trong những năm gần đây qua các hội nghị

chính về lĩnh vực này được tài trợ bởi ACM SIGIR [1] như SIGIR [2] và CIKM [3].

Mô hình không gian vectơ xem mỗi tài liệu và truy vấn là một vectơ đặc trưng và đánh giá độ liên quan giữa tài liệu và truy vấn dựa trên *độ tương tự* (*similarity*) giữa hai vectơ này. Tuy nhiên các đặc trưng là gì, trọng số (giá trị của mỗi đặc trưng) và độ tương tự được tính như thế nào không nằm trong mô hình. Các đặc trưng thường được chọn là các *từ đơn* (*word*), hoặc *cụm từ* (*phrase*). Một cách đánh giá trọng số rất thành công là "*pivoted normalization weighting*" của đại học Cornell [Singhal et al., 1996]. Một mở rộng của mô hình không gian vec tơ là mô hình *latent semantic indexing* (*LSI*) [Deerwester et al., 1990]. Mô hình này đưa ra cách thu giảm số chiều của *không gian đặc trưng* (*term space*) nhằm biểu diễn tốt hơn vectơ tài liệu và truy vấn.

Mô hình xác suất cổ điển binary independence retrieval (*BIR*) xác định *độ liên quan* giữa *tài liệu d* và *truy vấn q* bằng cách ước lượng "*xác suất mà người sử dụng sẽ tìm thấy tài liệu d thỏa mãn câu truy vấn q*".

Mô hình ngôn ngữ tìm cách "*ước lượng sự phân phối của các từ trong một ngôn ngữ*". Mỗi tài liệu được xem như một *mẫu ngẫu nhiên* (*random sample*) từ một *mô hình ngôn ngữ bên dưới. Các tài liệu được xếp hạng dựa vào khả năng mỗi mô hình ngôn ngữ của nó sinh ra các từ trong câu truy vấn.*

Mô hình mạng suy diễn đánh giá *độ liên quan* của *tài liệu* đến *truy vấn* dựa vào *khả năng suy ra* (*inferring/proving*) *truy vấn* từ *tài liệu. Độ liên quan* được tính dựa trên việc "*ước lượng xác suất truy vấn được thỏa mãn với điều kiện tài liệu được tìm thấy".*

[1] http://www.sigir.org/

[2] http://portal.acm.org/event.cfm?id=RE160

[3] http://portal.acm.org/event.cfm?id=RE302

4 Đánh giá hệ thống truy tìm thông tin

Để đánh giá một hệ thống truy tìm thông tin, hai tiêu chí chính được sử dụng là *mức độ chính xác của kết quả* (*effectiveness*) và *thời gian đáp trả của hệ thống* (*efficiency*). Để đánh giá mức độ chính xác của kết quả, *độ chính xác trung bình* (*average precision*) là một biện pháp phổ biến thường được dùng. Độ chính xác trung bình được tính dựa trên hai chỉ số: *độ bao phủ* (*recall*) và *độ chính xác* (*precision*). Chỉ số *recall* được tính bằng tỉ lệ giữa số tài liệu phù hợp mà hệ thống tìm được với số tài liệu phù hợp trong toàn bộ tập tài liệu cần tìm. Chỉ số *precision* được tính bằng tỉ lệ giữa số tài liệu phù hợp tìm được với số tài liệu tìm được. Thực tế cho thấy luôn có một *sự cân bằng* (*tradeoff*) giữa 2 chỉ số này, tức là khi ta tìm cách là cho chỉ số *recall* tăng lên thì chỉ số *precision* sẽ giảm xuống và ngược lại. *Độ chính xác trung bình* là một giá trị duy nhất được tính bằng cách xác định giá trị *precision* tại các *recall* khác nhau và lấy trung bình [Singhal, 2001].

Tuy nhiên, bài toán tìm kiếm thông tin là một bài toán không *được định nghĩa tốt* (*well-defined*) bởi vì chất lượng (khả năng trả về kết quả phù hợp với yêu cầu người dùng) của một hệ thống tìm kiếm chỉ có thể được đánh giá bởi *người dùng* (*user*) một cách chủ quan. Vì vậy, để xác định hệ thống tìm kiếm (mô hình, thuật toán) nào tốt hơn, chúng ta phải dựa vào những *đánh giá* (*judgments*) của người dùng, hoặc là tập dữ liệu dùng để đánh giá phải được tạo ra dựa trên những đánh giá của người dùng. Các kết quả tìm kiếm của một hệ thống tìm kiếm thường được đánh giá theo hai cách: (1) tiến hành nghiên cứu trên những người dùng sử dụng hệ thống để đánh giá chất lượng của quá trình tìm kiếm và kết quả (*user based evaluation*), hoặc (2) phát triển một *bộ sưu tập dữ liệu đánh giá chuẩn* (*standard test collection*) và thử nghiệm một hệ thống trên tập dữ liệu này để đánh giá chất lượng của các kết quả tìm kiếm (*system based evaluation*) [Zhai, 2008].

Cách thứ nhất cho phép chúng ta thấy được hiệu quả thực tế của một hệ thống.

Tuy nhiên, do sự tham gia của những người dùng khác nhau hoặc những trạng thái khác nhau của cùng một người dùng (một người dùng sử dụng một câu truy vấn cho hệ thống A và sau đó cũng dùng câu truy vấn đó cho hệ thống B sẽ có thể đánh giá không khách quan do đã quen với chủ đề sau khi sử dụng hệ thống A), rất khó để có thể so sánh hai hệ thống một cách đáng tin cậy bằng cách sử dụng phương pháp này.

Chính vì những lý do trên mà cách thứ hai là cách thường được chọn (cho đến hiện nay) trong việc đánh giá kết quả tìm kiếm của một hệ thống tìm kiếm (đặc biệt là trong nghiên cứu). Cách đánh giá này được đề xuất bởi Cleverdon và các cộng sự vào những năm 1960 [Cleverdon, 1967, 1991] và thường được gọi là *phương pháp đánh giá Cranfield* (*Cranfield evaluation method*). Theo phương pháp này, một bộ sưu tập các tài liệu và các truy vấn thực tế sẽ được chọn làm mẫu (thường là từ người sử dụng thực tế), sau đó, người sử dụng thực tế (lý tưởng nhất là những người thiết kế các truy vấn) sẽ đánh giá tất cả các tài liệu ứng với mỗi truy vấn để xác định những tài liệu có liên quan. Một bộ sưu tập đánh giá chuẩn gồm ba thành phần: (1) *bộ sưu tập tài liệu* (*document collection*), (2) *bộ sưu tập truy vấn* (*query collection*), và (3) các *đánh giá về độ liên* cho tất cả các truy vấn (*relevance judgments*). Một tài liệu được giả định là có liên quan hoặc không có liên quan đến truy vấn (*binary relevance*). Những đánh giá về sự liên quan cho tất cả các truy vấn sau khi thu thập sẽ được sử dụng để xác định tính chính xác của kết quả trả về (thường xếp hạng các các tài liệu liên quan đến câu truy vấn theo thứ tự giảm dần). Một xếp hạng lý tưởng sẽ đưa tất cả các tài liệu liên quan lên trên tất cả những tài liệu không liên quan. Phương pháp này được chấp nhận bởi *TREC* [4], một hội nghị được tổ chức hằng năm nhằm đánh giá kết quả của các kỹ thuật tìm kiếm.

[4] http://trec.nist.gov/

5 Tài liệu tham khảo

Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999. ISBN 0-201-39829-X.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005.

Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. Adapting ranking svm to document retrieval. In *SIGIR*, pages 186–193, 2006.

Cyril Cleverdon. The cranfield tests on index language devices. *Aslib Proceedings*, 19(6):173–194, 1967.

Cyril W. Cleverdon. The significance of the cranfield tests on index languages. In *SIGIR*, pages 3–12, 1991.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6): 391–407, 1990.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008. ISBN 978-0-521-86571-5.

Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, 1998.

S. Robertson and K. S. Jones. Relevance weighting of search terms. Number 27, 1976.

Gerard Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *SIGIR*, pages 21–29, 1996.

Howard R. Turtle and W. Bruce Croft. Inference networks for document retrieval. In *SIGIR*, pages 1–24, 1989.

ChengXiang Zhai. *Statistical Language Models for Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2008.