Làm giàu đặc trưng cho bài toán phân lớp truy vấn

Học viên: Nguyễn Thành Trung
Đơn vị công tác: Công ty CP CNTT, VT & TĐH Dầu khí
Email: trungnt1983@yahoo.com

GVHD: TS. Nguyễn Trí Thành
Đơn vị công tác: ĐH Công nghệ - ĐH Quốc gia Hà Nội
Email: ntthanh@vnu.edu.vn

Từ khóa: **chủ đề ẩn, LDA, máy tìm kiếm, phân lớp, truy vấn.**

1. GIỚI THIỆU BÀI TOÁN

Bài toán phân lớp truy vấn là một bài toán thuộc lĩnh vực tìm kiếm thông tin. Nội dung của bài toán là gán câu truy vấn của người sử dụng vào lớp đã được định nghĩa. Bài toán phân lớp truy vấn và bài toán phân lớp văn bản có nhiều đặc điểm giống nhau nhưng do các câu truy vấn rất ngắn và nhập nhằng nên bài toán này khó hơn rất nhiều so với bài toán phân lớp văn bản. Các thuật toán phân lớp truy vấn hiện nay đều chưa cho độ chính xác cao [1, 2, 5].

Bài toán phân lớp truy vấn có thể được ứng dụng trong các máy tìm kiếm. Nếu câu truy vấn đầu vào của người dùng được phân lớp thì máy tìm kiếm chỉ tìm trong lĩnh vực liên quan đến câu truy vấn đó, các kết quả trả về sẽ ít hơn và chính xác hơn. Ngoài ra bài toán phân lớp truy vấn còn được ứng dụng trong máy siêu tìm kiếm, quảng cáo trực tuyến.

Luận văn nghiên cứu bài toán phân lớp truy vấn và đề xuất một phương pháp làm giàu câu truy vấn để nâng cao hiệu quả của bộ phân lớp.

2. NỘI DUNG LUẬN VĂN

A. Mô hình phân tích chủ đề ẩn với LDA

LDA (Latent Dirichlet Allocation) là một mô hình sinh xác suất cho tập dữ liệu rời rạc dựa trên phân phối Dirichlet dựa trên ý tưởng: mỗi tài liệu là sự trộn lẫn của nhiều chủ đề, mỗi chủ đề là một phân phối xác suất trên tập các từ. Về bản chất, LDA là mô hình Bayesian ba mức: mức kho dữ liệu, mức tài liệu và mức từ [3].

Mô hình LDA rất giống với mô hình pLSA (probabilistic Latent Semantic Analysis) [4], chỉ có một điểm khác là mô hình LDA sử dụng phân phối Dirichlet để phân phối chủ đề.

B. Đề xuất mô hình làm giàu câu truy vấn

Ý tưởng của mô hình làm giàu câu truy vấn là dựa vào dụng các chủ đề ẩn được sinh ra trong mô hình phân tích chủ đề ẩn LDA. Nguồn sinh ra các tri thức mới là kho dữ liệu Internet thông qua máy tìm kiếm Google. Dựa vào các cách sử dụng máy tìm kiếm Google để lấy dữ liệu, tác giả đề xuất hai mô hình làm giàu câu truy vấn:

- Mô hình 1: Tìm kiếm trên Google các câu truy vấn trong tập dữ liệu.
- Mô hình 2: Tìm kiếm trên Google các câu truy vấn của người sử dụng.

Các bước thực hiện mô hình 1:

- Thực hiện ngoại tuyến: Các câu truy vấn trong tập dữ liệu được tìm kiếm trên Google, lấy các kết quả cao nhất sau đó tổng hợp kết quả lại và đưa vào mô hình LDA để sinh ra các chủ đề ẩn. Các chủ đề ẩn sau đó được lọc ra để lấy các chủ đề ẩn gần với các lớp nhất.
- Thực hiện trực tuyến: Câu truy vấn sau khi được tiền xử lý sẽ được tính độ tương tự với các chủ đề ẩn đã được lựa chọn để tìm độ tương tự lớn nhất, sau đó câu truy vấn được làm giàu bằng cách thêm vào từ có xác suất cao nhất của chủ đề ẩn.

Các bước thực hiện mô hình 2: Câu truy vấn của người sử dụng được tìm kiếm trên Google, lấy các kết quả cao nhất sau đó tổng hợp kết quả lại và đưa vào mô hình LDA để sinh ra các chủ đề ẩn. Các chủ đề ẩn sau đó được lọc ra để lấy các chủ đề ẩn gần với các lớp nhất. Câu truy vấn của người sử dụng sau khi được tiền xử lý sẽ được tính độ tương tự với các chủ đề ẩn đã được lựa chọn để tìm độ tương tự lớn nhất, sau đó câu truy vấn được làm giàu bằng cách thêm vào từ có xác suất cao nhất của chủ đề ẩn.

C. Thực nghiệm và đánh giá

Bộ dữ liệu được sử dụng trong quá trình thực nghiệm là truy vấn của trang AOL trong mùa thu năm 2004 [1, 2]. Quá trình thực nghiệm với cả hai mô hình cho thấy độ chính xác và độ đo F đều tăng so với kết quả ban đầu. Mô hình 2 có độ chính xác cao hơn nhưng thời gian thực hiện lâu hơn so với mô hình 1.

3. KẾT LUẬN

Quá trình thực nghiệm đã đạt kết quả khả quan cho thấy tính đúng đắn của việc lựa chọn phương pháp. Tuy độ chính xác của phân lớp tăng lên không cao nhưng hứa hẹn nhiều tiềm năng để phát triển.

TÀI LIỆU THAM KHẢO

[1] S. M. Beitzel et al. Improving Automatic Query Classification via Semi-supervised Learning. *The 5th IEEE International Conference on Data Mining*, 2005.

[2] S. M. Beitzel. *On Understanding and Classifying Web Queries*. PhD Thesis, Illinois Institute of Technology, 2006.

[3] D. Blei M. et al. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, Volume 3, pp. 993-1022.

[4] T. Hofmann. Probabilistic Latent Semantic Indexing, *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval*, pp. 50-57, 1999.

[5] D. Shen et al. Query enrichment for web-query classification. *Journal ACM Transactions on Information Systems*, Volume 24, Issue 3, pp. 320-352, 2006.