

NGHIÊN CỨU, PHÁT TRIỂN PHƯƠNG PHÁP TÍNH ĐỘ TƯƠNG TỰ CÂU TRUY VẤN TRONG HỆ TÌM KIẾM VÀ ỨNG DỤNG THỬ NGHIỆM VÀO MỘT HỆ TÌM KIẾM THỰC THỂ TIẾNG VIỆT

Học viên: Nguyễn Thị Thu Chung
Đơn vị công tác: TTPM - Tập đoàn Viettel
Email: thuchungnd@gmail.com

GVHD: PGS.TS. Hà Quang Thụy
Đơn vị công tác: Đại học Công nghệ - Đại học Quốc gia Hà Nội
Email: thuyhq@vnu.edu.vn

Từ khóa: Userlog, độ tương tự, truy vấn, chủ đề ẩn, WordNet, máy tìm kiếm

1. GIỚI THIỆU BÀI TOÁN

Tính độ tương tự truy vấn là bài toán con của bài toán tính độ tương tự của những đoạn văn bản rất ngắn. Nó có tính ứng dụng cao để giải quyết các vấn đề giúp việc tìm kiếm của người dùng được thuận tiện và chính xác hơn.

2. NỘI DUNG LUẬN VĂN

Luận văn trình bày 3 phương pháp tính độ tương tự truy vấn chính là Phương pháp thống kê, phương pháp sử dụng xử lý ngôn ngữ tự nhiên, phương pháp sử dụng kết quả trả về từ máy tìm kiếm. Với phương pháp thống kê, các độ đo thường được sử dụng để tính độ tương tự là độ đo cosin, Dice, Jaccard, overlap, độ đo kết hợp. Với phương pháp xử lý ngôn ngữ tự nhiên, luận văn trình bày hai phương pháp là phương pháp sử dụng Wordnet và phương pháp sử dụng Hidden topic sử dụng mô hình LDA. Phương pháp sử dụng kết quả trả về từ máy tìm kiếm là một phương pháp đặc trưng để tính độ tương tự giữa các truy vấn, nó sử dụng bộ dữ liệu userlog được lấy từ các máy chủ tìm kiếm, lưu lại lịch sử truy vấn của người dùng. Userlog có cấu trúc cơ bản gồm: lượt truy vấn, câu truy vấn, các liên kết người dùng chọn mở. Sử dụng dữ liệu này với công thức tính độ tương tự

$$\text{sim_result}(Q_i, Q_j) = \frac{N(R_{ij})}{\text{Max}(N(U(Q_i)), N(U(Q_j)))}$$

Trong đó:

- $\text{sim_result}(Q_i, Q_j)$: Độ tương tự giữa hai câu truy vấn Q_i và Q_j

- $N(U(Q_i))$: Số lượng liên kết tài liệu được người dùng nhấn vào khi thực hiện tìm kiếm với truy vấn Q_i

3. THỰC NGHIỆM

Áp dụng phương pháp tính độ tương tự truy vấn dựa vào kết quả trả về của máy tìm kiếm với công thức

$$\text{sim_result}(Q_i, Q_j) = \frac{\sum (x_i + x_j)}{\text{Max}(N(U(Q_i)), N(U(Q_j)))}$$

Trong đó: $\{x_1, x_2, \dots, x_n\}$ tương ứng là trọng số của liên kết 1, liên kết 2, ..., liên kết n

Áp dụng tính độ tương tự của 5000 được lấy từ máy chủ google mà công ty ITIM cung cấp, truy vấn trên máy tìm kiếm tiếng Việt đạt kết quả

STT	Truy vấn truy vấn	Độ tương tự
1	báo tin nhanh	0.9
2	tin tức online tin nhanh	0.45
3	nghe nhạc online mp3 trực tuyến	0.6166666
4	baby einstein nhạc cho bà bầu	0.46249998
5	báo gia đình nuôi dạy con	0.46249998

4. KẾT LUẬN

Luận văn đã tìm hiểu một số phương pháp tính độ tương tự giữa các câu truy vấn trong máy tìm kiếm. Sử dụng một phương pháp vào tính độ tương tự giữa các truy vấn bằng ngôn ngữ tiếng Việt và áp dụng vào máy tìm kiếm tiếng Việt