

NGHIÊN CỨU, CẢI TIẾN PHƯƠNG PHÁP MỞ RỘNG CÂU HỎI VÀ TÍCH HỢP VÀO HỆ THỐNG TÌM KIẾM THỰC THỂ TIẾNG VIỆT

Học viên: Phạm Thị Thu Uyên
Đơn vị công tác: Phòng Quản lý Chất lượng, Trung tâm phần mềm Viễn thông Viettel
Email: phamthuuyen@gmail.com

GVHD: PGS.TS Hà Quang Thụy
Đơn vị công tác: Bộ môn Hệ thống thông tin, Khoa Công nghệ Thông tin, Đại học Công nghệ, ĐHQGHN
Email: thuyhq@vnu.edu.vn

Từ khóa: Mở rộng truy vấn, query log, máy tìm kiếm, tiếng Việt.

1. GIỚI THIỆU BÀI TOÁN

Tìm kiếm thông tin là nhu cầu cần thiết trong cuộc sống, con người có được thông tin từ rất nhiều nguồn tài liệu khác nhau như sách vở, tạp chí,... Hiện nay, việc tin học hóa vào cuộc sống ngày càng cao và thời đại thông tin bùng nổ thì Internet là một trong những nguồn tài liệu mà con người thường sử dụng nhất. Internet là một kho dữ liệu đồ sộ, tuy nhiên để khai thác được nó là một thách thức lớn vì dữ liệu trên Internet quá đa dạng và không có cấu trúc. Vì thế, để tìm được thông tin cần thiết và chính xác với mong muốn của người sử dụng là một bài toán khó. Có nhiều phương pháp được nghiên cứu, luận văn thực hiện lựa chọn phương pháp xử lý câu hỏi hay truy vấn đầu vào, cụ thể là phương pháp Mở rộng truy vấn.

2. NỘI DUNG LUẬN VĂN

Luận văn thực hiện tìm hiểu và giới thiệu ba phương pháp chính được sử dụng cho việc Mở rộng truy vấn là: Phương pháp thủ công, phương pháp tự động, phương pháp kết hợp. Dựa vào các điều kiện hiện tại về kỹ thuật xử lý ngôn ngữ cũng như tài nguyên ngôn ngữ học, luận văn thực hiện lựa chọn phương pháp kết hợp và đề xuất một mô hình Mở rộng truy vấn sử dụng đồ thị khái niệm thông qua lịch sử giao dịch của người dùng. Đồng thời, áp dụng mô hình Mở rộng truy vấn để tích hợp vào hệ thống tìm kiếm thực thể tiếng Việt.

• Mô hình Mở rộng truy vấn:

- Bước 1: Xác định mối quan hệ giữa các câu truy vấn (bước thực hiện xử lý offline)
- Bước 2: Thực hiện xây dựng đồ thị khái niệm thể hiện mối quan hệ giữa các câu truy vấn
- Bước 3: Xác định khái niệm
- Bước 4: Thực hiện mở rộng truy vấn
- Bước 5: Cải tiến

Luận văn đã thực hiện thực nghiệm trên một số câu truy vấn như: “Máy tính”, “Trạng lương”, “Chung cư”. Đồng thời, luận văn thực hiện đánh giá kết quả đạt được bằng việc thực hiện so sánh kết quả và số lượng trả về của hệ thống giữa câu truy

vấn ban đầu và câu truy vấn mở rộng thông qua ba tiêu chí như sau:

- Số lượng trang web tìm được
- Số lượng trang web có nội dung khác nhau
- Số lượng trang web thực sự chứa cụm từ truy vấn.

3. KẾT LUẬN

- Trình bày một cách có hệ thống một số nội dung liên quan trực tiếp tới bài toán mở rộng truy vấn trong hệ thống tìm kiếm như Xử lý truy vấn đầu vào, Thu thập dữ liệu, Xếp hạng tài liệu để làm rõ ý nghĩa và nội dung của bài toán mở rộng truy vấn trong hệ thống tìm kiếm.
- Phân tích ba phương pháp chính mở rộng câu hỏi là Phương pháp thủ công, Phương pháp tự động, Phương pháp kết hợp. Qua xem xét điều kiện về bộ công cụ xử lý tiếng Việt và kho ngữ liệu tiếng Việt, luận văn định hướng vào phương pháp kết hợp để mở rộng truy vấn tiếng Việt.
- Đề nghị một mô hình Mở rộng truy vấn tiếng Việt dựa trên phương pháp sử dụng tập query log [4]. Thông qua đồ thị khái niệm được xây dựng từ tập các mối quan hệ giữa các câu truy vấn tiếng Việt, luận văn áp dụng mô hình đề nghị vào Hệ thống tìm kiếm tiếng Việt Google.
- Kết quả ban đầu của thực nghiệm cho thấy, tập câu truy vấn mở rộng chứa nhiều thông tin liên quan tới câu truy vấn, có thể hỗ trợ cho người dùng trong việc tìm kiếm chính xác được thông tin mong muốn. Đồng thời, kết quả trả về từ máy tìm kiếm khi nhập các câu truy vấn mở rộng có độ chính xác và số lượng kết quả cao hơn.

Do hạn chế về thời gian và kiến thức có sẵn, luận văn mới chỉ dừng lại ở mức thử nghiệm mô hình trên một số câu truy vấn và áp dụng kết quả đánh giá của bài toán Mở rộng truy vấn vào hệ thống tìm kiếm tiếng Việt. Trong thời gian tới, tiến hành thực nghiệm trên nhiều câu truy vấn có trong tập query log của Google, đồng thời thực hiện tích hợp module Mở rộng truy vấn vào một hệ thống tìm kiếm thực thể tiếng Việt nhằm đáp ứng nhu cầu tìm kiếm của người dùng.