THÔNG TIN LUẬN ÁN

Tên luận án:

NGHIÊN CỨU PHƯƠNG PHÁP XÂY DỰNG HỆ THỐNG QUẢN LÝ TÀI LIỆU VĂN BẢN DỰA TRÊN NGỮ NGHĨA

Chuyên ngành:	Khoa học máy tính
Mã số ngành:	62.48.01.01
Họ và tên Nghiên cứu sinh:	Huỳnh Thị Thanh Thương
Hướng dẫn khoa học:	PGS.TS. Đỗ Văn Nhơn
Cơ sở đào tạo:	Trường Đại học Công nghệ Thông tin - ĐHQG TP. HCM

1. TÓM TẮT

Luận án đề xuất một phương pháp mới cho bài toán Tìm kiếm tài liệu theo ngữ nghĩa thuộc một miền tri thức xác định, làm cơ sở khoa học cho việc thiết kế, xây dựng các hệ thống ứng dụng trong thực tiễn. Luận án nỗ lực nâng cao hiệu quả tìm kiếm thông qua việc nghiên cứu các phương pháp biểu diễn tài liệu cùng với kỹ thuật tính toán độ tương đồng ngữ nghĩa giữa tài liệu và câu truy vấn. Cách tiếp cận là biểu diễn văn bản dựa trên đồ thị keyphrase và tận dụng một ontology miền với độ mịn cao, được kiểm soát tốt để làm cơ sở cải thiện kết quả. Ngoài ra, luận án cũng tập trung vào việc nghiên cứu một giải pháp toàn diện cho việc thiết kế một loại hệ thống mới gọi là "Hệ quản lý cơ sở tài liệu văn bản theo ngữ nghĩa", thực hiện xây dựng một số hệ thống cụ thể để chứng minh tính hiệu quả và khả thi của các ý tưởng được đề xuất. Bên cạnh vấn đề tìm kiếm theo ngữ nghĩa, lợi ích của mô hình biểu diễn tài liệu dựa trên đồ thị và các kỹ thuật có liên quan còn được minh chứng thông qua bài toán Đo lường độ tương đồng ngữ nghĩa giữa hai tài liệu. Phương pháp mới tạo ra các biểu diễn có cấu trúc của văn bản bằng cách sử dụng những cơ sở tri thức có kích thức lớn và phổ biến như DBpedia, Wikipedia để thu thập thông tin chi tiết về các khái niệm, thực thể và các mối quan hệ ngữ nghĩa của chúng, do đó dẫn đến cách diễn giải "giàu tri thức" hơn cho tài liệu.

Các kết quả nghiên cứu được công bố trên các tạp chí và kỷ yếu hội nghị quốc tế chuyên ngành, được lập chỉ mục bởi các tổ chức có uy tín như Web of Science, Scopus, EI Compendex, Inspec, DBPL, ACM Digital Library, v.v.

2. CÁC ĐÓNG GÓP CHÍNH CỦA LUẬN ÁN

Luận án có các đóng góp chính như sau:

1) Đề xuất một phương pháp mới cho việc giải quyết bài toán Tìm kiếm tài liệu theo ngữ nghĩa thuộc một miền tri thức xác định, bao gồm: Một mô hình ontology CK-ONTO mô tả tri thức của lĩnh vực, làm căn cứ để biểu diễn ngữ nghĩa cho tài liệu; Các mô hình đồ thị keyphrase biểu diễn cho nội dung của tài liệu thuộc miền và kỹ thuật xây dựng đồ thị; Một kỹ thuật đo lường mức độ liên quan giữa tài liệu và câu truy vấn, dựa trên ý tưởng đánh giá độ tương đồng ngữ nghĩa giữa hai đồ thị keyphrase biểu diễn chúng.
2) Đề xuất một giải pháp tổng thể cho việc thiết kế và xây dựng một lớp hệ thống ứng dụng gọi là "Hệ quản lý cơ sở tài liệu văn bản theo ngữ nghĩa".

3) Xây dựng thử nghiệm 03 hệ thống ứng dụng: Hệ quản lý kho tài nguyên học tập về lĩnh vực Khoa học máy tính; Hệ thống hỗ trợ tìm kiếm việc làm và tuyển dụng ngành Công nghệ thông tin; Hệ thống hỗ trợ tìm kiếm, chọn lọc tin bài trên các báo mạng (lĩnh vực Lao động việc làm, Đầu tư công và đầu tư nước ngoài) phục vụ cho nhu cầu thực tế của Phòng Báo chí và Xuất bản của Sở Thông tin và Truyền thông Bình Dương.
4) Đề xuất một phương pháp mới giải quyết bài toán Đo lường độ tương đồng ngữ nghĩa giữa hai tài liệu.
5) Cơ sở tri thức của các lĩnh vực Khoa học máy tính, Việc làm ngành Công nghệ thông tin, Lao động việc làm, Đầu tư công và Đầu tư nước ngoài; Các bộ dữ liệu thử nghiệm phục vụ cho việc đánh giá hiệu quả của các hệ thống tìm kiếm tài liệu.

3. HƯỚNG PHÁT TRIỂN

Những vấn đề cần được tiếp tục nghiên cứu và phát triển bao gồm:

- Nghiên cứu các heuristics và cải tiến thuật toán để giảm độ phức tạp tính toán, tối ưu hóa hiệu suất của các giải thuật tìm kiếm.
- Phát triển phương pháp biểu diễn nội dung tài liệu theo hướng khái niệm, biểu diễn tri thức cho nhiều lĩnh vực có liên quan, trong đó vấn đề tích hợp tri thức cần được chú trọng.
- Thiết kế cơ chế cập nhật tự động ontology cũng như các thành phần khác bị ảnh hưởng bởi sự thay đổi (ví dụ như đồ thị keyphrase của các tài liệu); tăng cường khả năng suy luận trên ontology.
- Phát triển, mở rộng các phương pháp và kỹ thuật phù hợp cho ngôn ngữ tiếng Việt.
- Đa dạng hóa các thông tin quản lý, các yêu cầu tìm kiếm khác nhau, xử lý các truy vấn phức tạp bằng ngôn ngữ tự nhiên.
- Phát triển phương pháp lập chỉ mục tự động cho kho tài liệu, nghiên cứu sử dụng các cơ sở dữ liệu phân tán, cơ sở dữ liệu đồ thị, mô hình tính toán chuyên dùng trong việc xử lý dữ liệu đồ thị cực lớn, giúp tối ưu hóa quá trình tìm kiếm thông tin trong các kho dữ liệu lớn.
- Nghiên cứu phương pháp tích hợp mô hình biểu diễn tri thức và biểu diễn nội dung trong thiết kế "hệ truy vấn kiến thức và truy tìm tài liệu".

CÁN BỘ HƯỚNG DẪN	NGHIÊN CỨU SINH
Đỗ Văn Nhơn	**Huỳnh Thị Thanh Thương**