ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

HUỲNH THỊ THANH THƯƠNG

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: 62480101 (9480101)

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

TP HỒ CHÍ MINH - Năm 2023

Công trình được hoàn thành tại: **Trường Đại học Công nghệ Thông tin - Đại học Quốc gia TP.HCM**

Người hướng dẫn khoa học: **PGS.TS. Đỗ Văn Nhơn**

Phản biện độc lập 1: PGS.TS. Phạm Thế Bảo

Phản biện độc lập 2: PGS.TS. Nguyễn Thanh Bình

Luận án sẽ được bảo vệ trước

Hội đồng chấm luận án cấp Trường tại :

Trường Đại học Công nghệ Thông tin, Đại học Quốc gia TP. HCM, Khu phố 6, Phường Linh Trung, Thành phố Thủ Đức, Thành phố Hồ Chí Minh

vào lúc 08 giờ 30 ngày 17 tháng 01 năm 2024

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Thư viện ĐHQG-HCM
- Thư viện Trường Đại học Công nghệ Thông tin

Mục lục

Chương 1 GIỚI THIỆU TỔNG QUAN VỀ ĐỀ TÀI

Chương 1 giới thiệu tổng quan về đề tài bao gồm các khảo sát về thực trạng ứng dụng công nghệ thông tin trong công tác tổ chức lưu trữ, khai thác tìm kiếm tài liệu theo ngữ nghĩa, những phân tích đánh giá về thực trạng, nhu cầu và khả năng nghiên cứu phát triển giải pháp cũng như ứng dụng. Nội dung tiếp theo là giới thiệu về bài toán trọng tâm trong nghiên cứu, đó là Biểu diễn tài liệu văn bản và Tìm kiếm tài liệu trong kho lưu trữ thuộc một miền tri thức nhất định, tóm tắt tình hình nghiên cứu có liên quan, nêu lên những vấn đề còn tồn tại. Định hướng nghiên cứu, mục tiêu chung, nội dung phạm vi nghiên cứu và các đóng góp của luận án sẽ được trình bày ở phần cuối chương.

1.1 Thực trạng và nhu cầu xây dựng các hệ thống quản lý và tìm kiếm tài liệu theo ngữ nghĩa

Với lượng thông tin khổng lồ như hiện nay, nhu cầu tìm kiếm thông tin trong vô vàn các thông tin được lưu trữ là một yêu cầu hết sức cần thiết, đặc biệt nhu cầu quản lý tài liệu điện tử và thông tin khoa học công nghệ phục vụ chia sẻ tri thức ngày càng trở nên quan trọng. Nhìn chung, nguồn tài nguyên được lưu trữ dưới dạng dữ liệu văn bản là rất rộng lớn và giàu thông tin nhưng việc khai thác nguồn dữ liệu này vẫn chưa đạt hiệu quả cao. Sự gia tăng và bùng nổ của các cơ sở dữ liệu lớn làm cho việc tìm kiếm văn bản càng trở nên quan trọng hơn bao giờ hết. Chính vì vậy, việc nghiên cứu phương pháp quản lý và tìm kiếm tài liệu văn bản giúp cho người sử dụng có thể tìm kiếm được những thông tin cần thiết một cách chính xác, hiệu quả, phục vụ cho các mục đích trong công việc cũng như trong đời sống là rất cần thiết.

1.2 Tóm tắt tình hình nghiên cứu liên quan đến lĩnh vực của đề tài

Hiện nay, những hệ thống **tìm kiếm dựa trên khái niệm** (concept search, concept - based search) hay **tìm kiếm theo ngữ nghĩa** (semantic search) được nghiên cứu phát triển nhằm thay thế cho những hệ thống truyền thống vốn đã bộc lộ nhiều khuyết điểm lớn. Ý tưởng chính đằng sau các giải pháp tìm kiếm theo ngữ nghĩa là sử dụng các nguồn tài nguyên giàu ngữ nghĩa để giải nghĩa cho các từ/cụm từ, từ đó có thể giải nghĩa cho cả câu truy vấn và các tài liệu.

Ngày nay, vấn đề tìm kiếm tài liệu theo ngữ nghĩa phải đối mặt với nhiều thách thức lớn. Vấn đề quan trọng đầu tiên cần phải giải quyết là lựa chọn một phương pháp

biểu diễn cho tài liệu (document representation), tức là chuyển đổi tài liệu văn bản thành dạng có cấu trúc phù hợp với chương trình máy tính trong khi vẫn có thể mô tả được nội dung nồng cốt của văn bản đó. Với quan điểm rằng, hiểu nội dung của một tài liệu đòi hỏi phải có sự hiểu biết về các khái niệm, các thực thể chính trong tài liệu đó cũng như cách thức mà chúng liên hệ với nhau và hơn hết, đồ thị lại là một cấu trúc toán học có khả năng mô hình hóa mối quan hệ cùng với các thông tin quan trọng về cấu trúc một cách hiệu quả. Từ ý tưởng này, nhiều mô hình đồ thị đã được đề xuất như mạng ngữ nghĩa, đồ thị khái niệm CGs, CGs cải tiến, đồ thị hình sao, đồ thị tần số, đồ thị khoảng cách, đồ thị đồng hiện... được đánh giá là có nhiều tiềm năng sử dụng, có nền tảng lý thuyết chặt chẽ, rõ ràng và hiệu suất thực nghiệm tốt.

Ngoài ra, để rút trích khái niệm từ tài liệu, hệ thống cần sử dụng đến nguồn tri thức về lĩnh vực nhất định nào đó. Một số dạng nguồn tri thức có thể kể đến như cây khái niệm phân cấp (conceptual taxonomy), ontology miền (domain ontology), mạng ngữ nghĩa (semantic linguistic network of concept), từ điển đồng nghĩa (thesaurus). Trong số những mô hình này, có thể nói ngày nay ontology đang được chú ý nhiều nhất.

Hiện nay một trong những hướng nghiên cứu về ontology là cố gắng xây dựng các cơ sở lý thuyết và kỹ thuật tích hợp tri thức từ nhiều nguồn ontology khác nhau, cũng như việc xây dựng các mô hình ontology hướng đến chia sẻ và kết nối tri thức giữa nhiều hệ thống máy tính. Mục tiêu này đưa đến các cơ sở tri thức đa lĩnh vực cực kỳ đồ sộ như DBPedia, Yago, v.v... đều là những cơ sở tri thức uy tín và được sử dụng trong nhiều ứng dụng khác nhau. Tuy nhiên, ngay cả khi có sự trợ giúp của những nguồn tri thức đa lĩnh vực này, bài toán tìm kiếm trên thực tế vẫn là một thách thức lớn.

Vì thế trong lĩnh vực truy xuất thông tin hiện nay đang có xu hướng chuyển dịch sang việc tập trung vào các bài toán đặc thù trong một miền tri thức nhất định. Sự tập trung này cho phép ontology có thể được tùy biến phù hợp hơn với từng miền tri thức và từng bài toán cụ thể, qua đó giúp máy tính có thể hiểu chính xác hơn các tài liệu và câu truy vấn cần tìm kiếm. Đã có những ontology rất nổi tiếng và uy tín, được sử dụng trong nhiều nghiên cứu khác nhau như: ontology MeSH và SNOMED trong miền y khoa, PhySH miền vật lý, JEL trong miền kinh tế , AGROVOC và AgriOnt trong miền nông nghiệp, CSO trong miền Khoa học máy tính và MSC trong miền toán học, v.v... **Tuy đã có nhiều ontology được xây dựng và chia sẻ, hầu hết các ontology như vừa kể trên đều không được xây dựng để hướng đến bài toán truy xuất tài liệu nói chung, cũng như bài toán tìm kiếm tài liệu thuộc một miền tri thức nói riêng**.

1.3 Định hướng nghiên cứu và mục tiêu của luận án

1) Đề tài sẽ tập trung nghiên cứu 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 tìm kiếm tài liệu ứng dụng trong thực tiễn.

Bài toán tìm kiếm theo ngữ nghĩa trên một kho tài liệu D thuộc về một miền tri thức cụ thể $\mathbb{K}$, giới hạn trong phạm vi ngôn ngữ là tiếng Anh, được mô tả như sau: từ câu truy vấn người dùng nhập vào, hệ thống tìm kiếm và trả về danh sách các tài liệu (được sắp hạng) có nội dung liên quan và phù hợp với thông tin truy vấn. Những tài liệu này không nhất thiết phải chứa chính xác từ khóa tìm kiếm. Câu truy vấn là một phát biểu ở thể khẳng định (không phải là dạng câu hỏi đáp) bằng ngôn ngữ tự nhiên, được đặc tả dưới dạng gồm một hay nhiều từ (cụm từ) được phân cách với nhau bằng khoảng trắng, tối đa 10 cụm từ.

Đề tài sẽ nỗ lực cải thiện hiệu quả của việc tìm kiếm thông qua việc **nghiên cứu các phương pháp biểu diễn cho tài liệu văn bản cùng với kỹ thuật tính toán độ tương đồng ngữ nghĩa giữa tài liệu và câu truy vấn**. Phương pháp tiếp cận là dựa trên ontology và biểu diễn văn bản bằng đồ thị. Như vậy, với cách tiếp cận được nêu trên, các bài toán con cần giải quyết bao gồm:

a. Nghiên cứu mô hình ontology biểu diễn tri thức thuộc một miền tri thức nhất định, qua đó làm căn cứ để biểu diễn ngữ nghĩa cho tài liệu

b. Nghiên cứu mô hình và kỹ thuật biểu diễn (nội dung) tài liệu (trên cơ sở đã mô hình hóa được miền tri thức mà tài liệu thuộc về).

c. Tính khoảng cách ngữ nghĩa giữa các keyphrase (hay các khái niệm) thông qua việc khai thác nguồn tri thức ontology miền dựng sẵn

d. So khớp và tính toán mức độ tương đồng ngữ nghĩa giữa các cấu trúc biểu diễn cho nội dung của tài liệu và câu truy vấn

2) Nghiên cứu giải pháp thiết kế, xây dựng một lớp hệ thống mới, gọi là "Hệ thống quản lý cơ sở tài liệu văn bản theo ngữ nghĩa" và xây dựng một số hệ thống ứng dụng cụ thể để chứng minh tính hữu ích và khả thi của các ý tưởng nghiên cứu đã đề xuất.

3) Nghiên cứu một phương pháp mới cho bài toán Đo lường mức độ tương đồng ngữ nghĩa giữa hai tài liệu thuộc về một miền tri thức đặc biệt hoặc thuộc tri thức tổng quát nói chung. Trong đề tài này, 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.

1.4 Đóng góp 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. Kết quả này được công bố trong công trình [CT1][CT2][CT3] và đóng góp ở Chương 3 và Chương 4 của luận án.

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". Kết quả này được công bố trong các công trình [CT1][CT2][CT3], đóng góp ở Chương 5 của luận án.

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. Kết quả này được công bố trong các công trình [CT4][CT5], đóng góp ở Chương 6 của luận án.

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.

1.5 Kết chương

Chương 1 giới thiệu tổng quan về tình hình ứng dụng công nghệ thông tin trong công tác tổ chức lưu trữ, khai thác, tìm kiếm tài liệu theo ngữ nghĩa, phân tích đánh giá thực trạng, nhu cầu và khả năng nghiên cứu phát triển giải pháp cũng như ứng dụng. Với thách thức đặt ra về việc cải thiện hiệu quả trong tìm kiếm thông qua cải tiến độ chính xác và độ bao phủ, đề tài nghiên cứu các phương pháp biểu diễn ngữ nghĩa cho tài liệu cùng với kỹ thuật tính toán độ tương đồng ngữ nghĩa giữa tài liệu và câu truy vấn. Nhận thấy tiềm năng ứng dụng của cách tiếp cận dựa trên ontology và biểu diễn văn bản bằng đồ thị, đề tài sẽ tập trung phân tích khả năng ứng dụng của các mô hình, nghiên cứu các phương pháp và kỹ thuật đã có, qua đó tìm cách vận dụng, phối hợp, cải tiến, phát triển nhằm tăng cường hiệu quả giải quyết các bài toán đã đặt ra.

Chương 2 CƠ SỞ LÝ THUYẾT

Chương 2 trình bày cơ sở lý thuyết của đề tài liên quan đến vấn đề Biểu diễn tài liệu và Tìm kiếm tài liệu theo ngữ nghĩa, hệ thống hóa các công trình nghiên cứu trong và ngoài nước liên quan đến nội dung nghiên cứu của đề tài, từ đó, chỉ ra hướng tiếp cận và phương pháp làm nền tảng cho các đóng góp của luận án ở các chương sau.

2.1 Vấn đề tìm kiếm tài liệu theo ngữ nghĩa và các hướng tiếp cận

Tìm kiếm theo ngữ nghĩa là một hình thức tìm kiếm mà sử dụng "ngữ nghĩa tường minh" để giải quyết các nhiệm vụ cốt lõi trong tìm kiếm, nghĩa là sử dụng ngữ nghĩa để giải nghĩa cho câu truy vấn và tài liệu, so khớp câu truy vấn với tài liệu, đánh giá mức độ liên quan và xếp hạng kết quả trả về. Các giải pháp tìm kiếm theo ngữ nghĩa khác nhau về:

- Khía cạnh của dữ liệu (loại dữ liệu được quan tâm trong nghiên cứu là semantic data, semantic metadata hay raw data).
- Nhu cầu thông tin của người dùng (information needs).
- Mô hình biểu diễn tài liệu và câu truy vấn.
- Phương pháp tìm kiếm (còn gọi là truy hồi) thông tin, tài liệu.
- Mô hình ngữ nghĩa và nguồn tài nguyên ngữ nghĩa được sử dụng (semantic resource).
- Cấu trúc và cách thức xây dựng ontology (ontology structure), công nghệ ontology(ontology technology).
- Cách giải quyết các bài toán con trong tìm kiếm bao gồm biểu diễn và xử lý nội dung (ngữ nghĩa) câu truy vấn và tài liệu, bài toán so khớp và xếp hạng, bài toán rút trích các đơn vị thông tin như từ khóa (keyword), cụm từ khóa (keyphrase), khái niệm (concept), thực thể (entity) và mối quan hệ (relationship) từ tài liệu.

2.2 Vấn đề biểu diễn tri thức và các mô hình ngữ nghĩa

Nhìn chung, ta có thể phân loại các mô hình ngữ nghĩa theo hai nhóm: mô hình từ vựng (lexical model) và mô hình tri thức (nnowledge model). Điểm khác biệt rõ ràng nhất giữa hai nhóm mô hình này là ở chỗ: các mô hình từ vựng thể hiện ngữ nghĩa ở mức độ từ ngữ thông qua quan hệ giữa các từ, một mức độ thấp hơn nhiều so với mức độ khái niệm và thực thể ngoài đời thực như các mô hình tri thức. Phần lớn các

mô hình tri thức được xây dựng trên ba thành phần cơ bản: lớp (khái niệm) của các thực thể, mối quan hệ giữa các lớp (hay các thực thể) và thuộc tính của các thực thể.

Mục tiêu nghiên cứu được đặt ra trong đề tài là tập trung vào bài toán tìm kiếm trong một miền tri thức nhất định. Sự tập trung này đòi hỏi sử dụng đến các ontology miền như là một cơ sở ngữ nghĩa nhằm giảm thiểu sự tối nghĩa, sự nhập nhằng về nghĩa, qua đó giúp máy tính có thể hiểu chính xác hơn các tài liệu và câu truy vấn cần tìm kiếm. Đã có những ontology miền rất nổi tiếng và uy tín, được sử dụng trong nhiều nghiên cứu khác nhau.Tuy nhiên, hầu hết các ontology như vừa kể trên đều không được xây dựng để hướng đến bài toán Truy xuất tài liệu (Adhoc document retrieval). Khi khảo sát kỹ các ontology, chúng ta sẽ thấy sự khác nhau rõ rệt giữa chúng, ngay cả khi chúng đã được xây dựng cho những mục đích rất tương tự. Các ontology mô tả tri thức ở các mức độ chi tiết khác nhau và không có một khuôn dạng chung để biểu diễn thông tin liên quan giữa các ontology, dẫn tới khó có thể sử dụng lại ontology đã có trong một ứng dụng tìm kiếm mới mà đề tài đang hướng tới.

2.3 Vấn đề biểu diễn tài liệu văn bản

Mô hình biểu diễn văn bản truyền thống như mô hình túi từ (Bag of words), mô hình không gian vector (Vector Space Model) là các mô hình đơn giản và được sử dụng phổ biến nhất trong phần lớn các bài toán xử lý dữ liệu văn bản. Tuy nhiên, những mô hình truyền thống này lại tồn tại trong nó những hạn chế lớn mà chủ yếu là do sự yếu kém trong vấn đề biểu diễn thông tin.

Nhìn chung, dạng biểu diễn văn bản bằng vector có tốc độ tính toán nhanh, đặc biệt là có sẵn các thư viện tính toán được hỗ trợ từ các ngôn ngữ lập trình cấp cao. Tuy nhiên, hầu hết các kỹ thuật chủ yếu dựa trên thông tin về tần suất xuất hiện của từ, thiếu sự phản ánh về ngữ nghĩa của văn bản: bỏ qua các thông tin cấu trúc quan trọng như thứ tự sắp xếp các từ trong câu, vùng lân cận của từ, vị trí xuất hiện của từ trong văn bản, cấu trúc của một câu/đoạn văn, tính đồng xuất hiện của các từ trong một câu và đặc biệt nghĩa của từ cũng như mối quan hệ về ngữ nghĩa giữa các từ không được xét đến, cuối cùng là hạn chế của kỹ thuật rút trích đặc trưng. Bên cạnh đó, các phép biểu diễn có thể khó diễn nghĩa, tức là khó diễn dịch, giải thích hay thuyết minh bởi người đọc. Các kết quả có thể được chứng minh ở cấp độ toán học, nhưng khó có thể hiểu được trong ngôn ngữ tự nhiên. Một hình thức biểu diễn được xem là tốt khi mà người đọc có thể dễ dàng nắm bắt ý nghĩa của chúng và hiểu được kết quả trả về của hệ thống cũng như cách thức hệ thống trả về được những kết quả này.

Trong những năm gần đây, các phương pháp mô hình hóa văn bản thành đồ thị đang ngày càng được chú ý. Nhiều mô hình đồ thị không ngừng được nghiên cứu phát triển và được ứng dụng vào dãy rộng các bài toán liên quan đến xử lý văn bản và đây cũng là cách tiếp cận được lựa chọn trong đề tài này.

2.4 Những bài toán con trong nghiên cứu

Xây dựng một hệ thống quản lý và tìm kiếm tài liệu là cả một quá trình phức tạp bao gồm nhiều giai đoạn phát triển mà tìm kiếm chỉ là một trong các giai đoạn đó. Việc thiết kế, cài đặt hệ tìm kiếm tài liệu nói chung và Tìm kiếm theo ngữ nghĩa nói riêng đặt ra nhiều vấn đề cần giải quyết. Các vấn đề được quan tâm hàng đầu, thường được nhắc đến và cũng là mục đích của các nghiên cứu liên quan đến Hệ thống tìm kiếm tài liệu bao gồm:

- Phân tích nhu cầu thông tin của người dùng và Tương tác người dùng (User' Information needs, User interaction)
- Rút trích thông tin tự động (Information Extraction) liên quan đến việc rút trích các từ/cụm từ hay khái niệm đặc trưng cho tài liệu và câu truy vấn, rút trích quan hệ giữa các từ và khái niệm được đề cập đến trong văn bản.
- Biểu diễn tài liệu và lập chỉ mục ngữ nghĩa tự động (Document Representation, Semantic Indexing): biểu diễn nội dung (ngữ nghĩa) tài liệu tức là chuyển đổi tài liệu văn bản thành dạng có cấu trúc phù hợp với chương trình máy tính mà vẫn có thể mô tả được nội dung nồng cốt của văn bản đó; cùng với đó là vấn đề tổ chức lưu trữ kho tài liệu có kích thước lớn sao cho hỗ trợ tìm kiếm một cách hiệu quả.
- Đo lường mức độ tương đồng ngữ nghĩa giữa các từ hay khái niệm
- Tìm kiếm và sắp hạng hiệu quả (Effective ranking): đo lường sự liên quan giữa tài liệu và câu truy vấn và sắp hạng kết quả trả về.
- Đánh giá hiệu quả truy tìm (Evaluation, Testing and Measuring)

  Ngoài ra còn có các nghiên cứu về trực quan hóa văn bản, phát hiện xu thế, khám phá chủ đề, phân loại, gom cụm, tóm tắt văn bản, ...nhưng không phải là mục tiêu chính của đề tài này.

2.5 Kết chương

Chương 2 trình bày cơ sở lý thuyết nền tảng liên quan đến vấn đề Tìm kiếm tài liệu theo ngữ nghĩa, trong đó, trọng tâm là tìm hiểu, phân loại và phân tích ưu nhược điểm của các phương pháp tìm kiếm tài liệu đã có. Các phương pháp, kỹ thuật dựa trên ontology và đồ thị sẽ là cơ sở quan trọng cho việc nghiên cứu giải pháp đặt ra trong mục tiêu của đề tài.

Chương 3 CK-ONTO: MỘT MÔ HÌNH ONTOLOGY MIỀN CHO CÁC HỆ THỐNG TÌM KIẾM TÀI LIỆU THEO NGỮ NGHĨA

Chương 3 sẽ trình bày một đóng góp của luận án cho vấn đề biểu diễn tri thức thuộc một miền nhất định theo tiếp cận ontology, qua đó làm căn cứ để biểu diễn ngữ nghĩa cho tài liệu. Kết quả nghiên cứu được công bố trong các công trình [CT1], [CT2] của tác giả.

3.1 Giới thiệu

Một mô hình ontology mới được đề xuất dùng để biểu diễn tri thức về một lĩnh vực đặc biệt. Mô hình mới có sự cải tiến bằng cách chi tiết hóa cũng như chuẩn hóa một số thành phần, định nghĩa và bổ sung những thành tố mới, từ đó có thêm được nhiều thông tin cần thiết phù hợp hơn với tác vụ tìm kiếm tài liệu theo ngữ nghĩa. Việc chi tiết hóa giúp mô hình trở nên rõ ràng và tường minh hơn, trong khi chuẩn hóa giúp đảm bảo tính nhất quán và sự đồng nhất trong việc biểu diễn tri thức. Bên cạnh đó, việc bổ sung các thành tố mới cung cấp thêm thông tin quan trọng và đa dạng, nâng cao khả năng tìm kiếm và truy xuất tài liệu theo yêu cầu.

3.2 Mô hình Classed Keyphrase based Ontology

Định nghĩa 1 *Classed Keyphrase based Ontology (CK-ONTO) là một mô hình của tri thức miền, cung cấp một cơ sở ngữ nghĩa tường minh nhằm hỗ trợ giải quyết các nhiệm vụ cốt lõi trong tìm kiếm, bao gồm bốn thành phần như sau:*

(K, C, R, Rules), trong đó,

- K là tập hợp các keyphrase trong một miền tri thức nhất định
- C là tập hợp các lớp tương ứng với các khái niệm trong miền
- R là tập hợp các quan hệ giữa các keyphrase trong K và giữa những khái niệm trong C.
- Rules là tập hợp các luật suy diễn

Keyphrase có thể là một thuật ngữ biểu thị một khái niệm khoa học. Ngoài ra, keyphrase cũng có thể chỉ đến một thực thể có tên duy nhất trong lĩnh vực. Ví dụ, một số keyphrase trong lĩnh vực *Khoa học máy tính* có thể kể đến như *document modeling, demantic network, domain ontology, named entity, DBpedia, TREC, Alan Turing.*

Các lớp, tương ứng với các khái niệm, là trung tâm của hầu hết các ontology. Cấu trúc (định nghĩa) của mỗi khái niệm $c \in C$ có thể được mô hình hóa bởi một bộ gồm năm thành phần sau: ***(cnames, Statement, Kbs, Attrs, Insts)***, trong đó:

- $\emptyset \neq$ cnames $\subseteq K_C$ là tập hợp các keyphrase có thể được dùng gọi tên cho khái niệm này. Một cnames còn được gọi là một synset, nghĩa là một tập các từ đồng nghĩa có thể thay thế được cho nhau khi mô tả khái niệm.

- *Statement* là định nghĩa phi hình thức của khái niệm, giúp mô tả và xác định khái niệm dưới dạng ngôn ngữ tự nhiên.

- Kbs $\subseteq$ K là một tập các "keyphrase nền". Keyphrase nền là những keyphrase hình thành nên định nghĩa của khái niệm ở dạng ngôn ngữ tự nhiên, tức là những keyphrase đặc trưng dùng để mô tả khái niệm.

- *Attrs* hoặc là một tập rỗng hoặc là tập hợp các thuộc tính của lớp, đặc tả cấu trúc bên trong của lớp.

- Cuối cùng, *Insts* là một tập rỗng hoặc là một tập các thực thể, còn gọi là thể hiện, cá thể hay đối tượng của lớp.

Phần lớn sức mạnh của ontology nằm ở khả năng diễn đạt quan hệ. Tập hợp các quan hệ cùng nhau mô tả ngữ nghĩa của một lĩnh vực. Các quan hệ trong ontology được phân làm các nhóm: quan hệ giữa các lớp và quan hệ trực tiếp giữa các keyphrase. Thành phần R trong mô hình là một bộ gồm hai thành phần con $R = (R_{KK}, R_{CC})$, trong đó, R_{CC} là một tập hợp các quan hệ nhị phân trên C và R_{KK} là tập hợp các quan hệ nhị phân trên K.

Tùy thuộc vào miền tri thức, ta có nhiều quan hệ ngữ nghĩa (semantic relation) khác nhau trên khái niệm. Các quan hệ này có thể được chia thành hai nhóm chính: các quan hệ phân cấp (hierarchical relations) và các quan hệ không phân cấp (non-hierarchical relations). Bên cạnh đó, các quan hệ trên keyphrase cũng được chia thành ba nhóm chính: các quan hệ tương đồng, các quan hệ tổ hợp, các quan hệ ngữ nghĩa phát sinh từ những quan hệ giữa các khái niệm.

Rules là tập hợp các luật suy diễn trên các sự kiện liên quan đến keyphrase và khái niệm. Mỗi luật cho ta một qui tắc suy luận để từ các sự kiện đã biết suy ra được các sự kiện mới. Một luật có thể được mô hình hóa dưới dạng:

Việc bổ sung thành phần luật mang lại những lợi ích sau: 1) tiết kiệm được nhiều chi phí cho việc tạo lập và lưu trữ các quan hệ của ontology trên máy tính; 2) kiểm soát được mâu thuẫn trên dữ liệu, tránh được nhiều trường hợp thiếu sót từ việc tạo quan hệ bằng phương pháp thủ công với số lượng quan hệ lớn; 3) nhờ vào tập luật và thuật toán suy diễn trên tập luật mà ta có thể xác được mối quan hệ giữa hai keyphrase, giữa keyphrase và lớp, và giữa hai lớp mà không cần phải khảo sát và xây dựng quan hệ ngay từ ban đầu.

3.3 Vai trò của CK-ONTO trong giải pháp thiết kế các hệ thống tìm kiếm tài liệu

Vai trò quan trọng của ontology trong giải pháp là cung cấp một "cơ sở ngữ nghĩa tường minh" nhằm hỗ trợ giải quyết các nhiệm vụ cốt lõi trong tìm kiếm, bao gồm: 1)Rút trích tự động các keyphrase đặc trưng ngữ nghĩa của tài liệu cũng như các mối quan hệ giữa chúng; 2) Đo lường mức độ tương đồng ngữ nghĩa giữa các keyphrase hay khái niệm; 3) Các kỹ thuật khác hỗ trợ quá trình tìm kiếm như thay đổi cách đánh trọng số, mở rộng câu truy vấn, mở rộng tài liệu.

3.4 Xây dựng ontology miền theo mô hình CK-ONTO

Xây dựng một cơ sở tri thức phụ thuộc miền theo mô hình CK-ONTO là công việc đòi hỏi sự giám sát chặt chẽ của các chuyên gia trong lĩnh vực. Một nhóm các chuyên gia và kỹ sư tri thức sẽ chịu trách nhiệm xây dựng và cải tiến lược đồ ontology này. Quy trình xây dựng ontology miền thường phải trải qua các bước chính như: 1) Thu thập dữ liệu và tích hợp ontology; 2) Làm giàu ontology từ các tài liệu trên Web, từ tập văn bản; 3) Ánh xạ tới các nguồn tài nguyên hiện có; 4) Chuẩn hóa ontology.

Với ba hệ thống ứng dụng được nghiên cứu phát triển như đã kể trên, đề tài đã xây dựng các cơ sở tri thức tương ứng, được tổ chức theo mô hình CK-ONTO.

Vì tính hiệu quả trong việc giải quyết bài toán tìm kiếm tài liệu phụ thuộc rất lớn vào chất lượng của ontology, việc điều chỉnh lại ontology trong và sau quy trình xây dựng là không thể tránh khỏi, cần có sự can thiệp và giám sát của đội ngũ chuyên gia trong lĩnh vực. Đề tài đã xây dựng một giao diện web cho phép quản lý và điều phối công việc giữa các chuyên gia và kỹ sư tri thức có tham gia xây dựng hệ thống.

3.5 Kết chương

Chương 3 đề xuất một mô hình ontology CK-ONTO biểu diễn tri thức thuộc một miền tri thức nhất định, làm căn cứ để biểu diễn ngữ nghĩa cho tài liệu. Kỹ thuật xây dựng ontology cho lĩnh vực, cùng với cách thức tổ chức lưu trữ và quản lý ontology trên máy tính sẽ được trình bày trong Phụ lục 3 và Phụ lục 5 của luận án.

Vai trò quan trọng của ontology trong giải pháp là cung cấp một cơ sở ngữ nghĩa tường minh nhằm hỗ trợ giải quyết các nhiệm vụ cốt lõi trong tìm kiếm. Cấu trúc của ontology được thiết kế có tính tổng quát và dễ dàng mở rộng cho nhiều lĩnh vực khác nhau cũng như các loại hình ứng dụng khác nhau.

Chương 4 BIỂU DIỄN TÀI LIỆU DỰA TRÊN ĐỒ THỊ KEYPHRASE VÀ ĐÁNH GIÁ ĐỘ TƯƠNG ĐỒNG NGỮ NGHĨA TRONG TÌM KIẾM

Chương 4 đã trình bày một đóng góp quan trọng của luận án về một phương pháp mới cho việc giải quyết bài toán Tìm kiếm theo ngữ nghĩa trên một kho tài liệu văn bản có liên quan đến một miền tri thức nhất định. Hiệu quả tìm kiếm được cải thiện thông qua việc nghiên cứu các phương pháp biểu diễn ngữ nghĩa cho tài liệu văn bản theo tiếp cận đồ thị, cùng với kỹ thuật tính toán độ tương đồng ngữ nghĩa giữa tài liệu và câu truy vấn. Các kết quả được công bố trong công trình [CT1], [CT2] của tác giả.

4.1 Giới thiệu

Định nghĩa 2 *(Tài liệu được truy hồi) Cho trước T là tập hợp các đối tượng văn bản thuộc về một miền tri thức cụ thể* $\mathbb{K}$*, một tập tài liệu văn bản (thô)* $D = \{d_1, d_2, ...d_n\} \subseteq T$*, một tập các câu truy vấn mẫu* $Q = \{q_1, q_2, ..., q_m\} \subseteq T$*, một hàm* $\sigma : Q \times D \rightarrow [0, 1]$ *cho phép đánh giá độ tương đồng ngữ nghĩa giữa một câu truy vấn trong Q và một tài liệu trong D.*

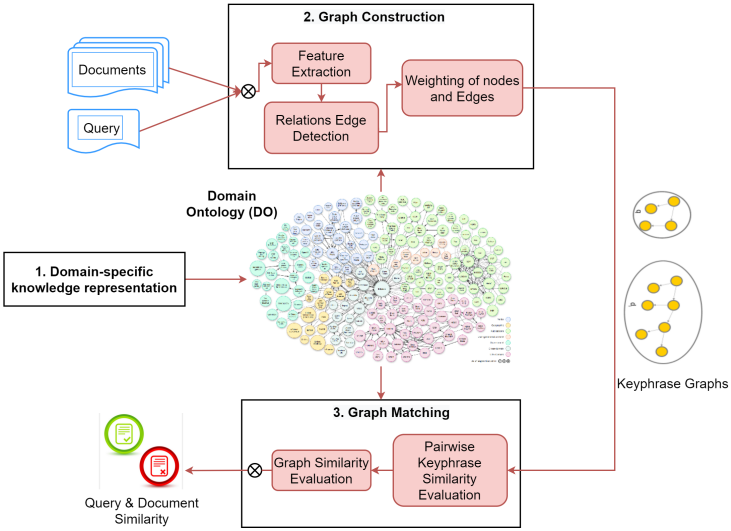
Cho $q \in Q$ *là một câu truy vấn và* $\tau \in \mathbb{R}$ *là một giá trị ngưỡng mong muốn. Tập các tài liệu trong* D *được truy hồi phù hợp với* q*, ký hiệu* $\mathrm{Rel}(q, D)$*, được định nghĩa như sau:*

$\mathrm{Rel}(q, D) = \{d \in D | \sigma(q, d) > \tau\}$

Tìm kiếm tài liệu cả một quá trình phức tạp bao gồm nhiều công đoạn xử lý và đặt ra nhiều vấn đề cần giải quyết. Hình 4.1 cung cấp một cách nhìn tổng quan về cách tiếp cận của đề tài cho bài toán này.

4.2 Biểu diễn tài liệu văn bản

Sau khi hệ thống hóa các mô hình biểu diễn văn bản bằng đồ thị đã có, phân tích khả năng ứng dụng và hạn chế của từng loại mô hình, đề tài đề xuất một nhóm mô hình đồ thị mới phù hợp hơn cho các nhiệm vụ tìm kiếm văn bản, gọi chung là **Đồ thị keyphrase (Keyphrase Graph)**. Đề tài sử dụng thuật ngữ "Keyphrase Graphs" (viết tắt là KGs) để chỉ đến một nhóm các mô hình và sử dụng các thuật ngữ đặc biệt



Hình 4.1 Quy trình tính toán độ tương đồng giữa tài liệu và câu truy vấn

như "simple keyphrase graph", "weighted keyphrase graph", "full weighted keyphrase graph" khi định nghĩa về mặt toán học cho những mô hình này.

Định nghĩa 3 *(Đồ thị keyphrase đơn giản)*

Định nghĩa 4 *(Đồ thị keyphrase có trọng số)*

Định nghĩa 5 *(Đồ thị keyphrase con)*

Định nghĩa 6 *(Đồ thị keyphrase con có trọng số)*

Định nghĩa 7 *(Đồ thị keyphrase có trọng số biểu diễn cho tài liệu)*

Định nghĩa 8 *(Đồ thị keyphrase có trọng số đầy đủ)*

Cho một tập tài liệu văn bản $\mathrm{D} = \{\mathrm{d}_1, \mathrm{d}_2, ...\mathrm{d}_\mathrm{n}\}$ *thuộc về một miền tri thức đặc biệt* $\mathbb{K}$, $\mathrm{O} = (\mathrm{K}, \mathrm{R}_{\mathrm{KK}})$ *là một mô hình con đạt được từ CK-ONTO biểu diễn* $\mathbb{K}$, *gọi* fKGs *là tập các đồ thị keyphrase có trọng số đầy đủ. Một hàm* $\mathrm{fulldocKG} : \mathrm{D} \rightarrow \mathrm{fKGs}$ *cho tương ứng mỗi tài liệu* $\mathrm{d} \in \mathrm{D}$ *một biểu diễn ngữ nghĩa dưới dạng đồ thị keyphrase* $\mathrm{fulldocKG}(\mathrm{d}) \in \mathrm{fKGs}$.

Một đồ thị keyphrase có trọng số đầy đủ (full weighted keyphrase graph), biểu diễn cho tài liệu d, ký hiệu là $\mathrm{fulldocKG}(\mathrm{d})$, *được định nghĩa trên* O, *là một bộ*

$(\mathrm{V}^{\mathrm{d}}, \mathrm{E}_1^{\mathrm{d}}, \mathrm{E}_2^{\mathrm{d}}, \phi_1^{\mathrm{d}}, \phi_2^{\mathrm{d}}, \mathrm{l}_{\mathrm{E}_1}^{\mathrm{d}}, \mathrm{l}_{\mathrm{E}_2}^{\mathrm{d}}, \mathrm{w}_{\mathrm{V}}^{\mathrm{d}}, \mathrm{w}_{\mathrm{E}}^{\mathrm{d}})$

thỏa mãn những điều kiện sau:

- $(V^d, E_1^d, \phi_1^d, l_{E_1}^d, w_V^d, w_E^d)$ là một đồ thị keyphrase có trọng số biểu diễn d .

- E_2^d là tập các cạnh có hướng biểu diễn những mối quan hệ về cú pháp (hay cấu trúc) giữa những đỉnh keyphrase (tập cạnh của đồ thị là $E^d = E_1^d \cup E_2^d$) và $\phi_2^d : E_2^d \rightarrow \{(x,y)|(x,y) \in {V^d}^2, x \neq y\}$ ánh xạ mỗi cạnh đến một cặp có thứ tự của hai đỉnh phân biệt. Ngoài những mối quan hệ ngữ nghĩa, hai đỉnh keyphrase $k_1, k_2 \in V^d$ cũng có thể được kết nối với nhau nếu có tồn tại một hình thức nào đó thể hiện mối quan hệ cú pháp giữa chúng chẳng hạn như quan hệ đồng xuất hiện hoặc quan hệ ngữ pháp trong câu.

- $l_{E_2}^d : E_2^d \rightarrow T_S$ là hàm gán nhãn cho những cạnh trong E_2^d. T_S là tập tên (hay ký hiệu) của những quan hệ về cú pháp được dùng để gán nhãn cho các cạnh này.

- $w_E^d : E^d \rightarrow \mathbb{R}^+$ được sử dụng cho việc đánh trong số các cạnh. Các trọng số nắm bắt mức độ liên quan nhau giữa các keyphrase trong đồ thị.

- Hai keyphrase được kết nối với nhau bởi quan hệ đồng hiện nếu chúng có xuất hiện cùng với nhau trong một câu của tài liệu. Khi đó, cạnh liên kết chúng được gán nhãn bởi "co-occurrence", hướng của cạnh này được xác định dựa vào thứ tự xuất hiện của hai keyphrase. Trọng số của cạnh phản ánh hai keyphrase có liên quan với nhau mạnh mẽ như thế nào và có thể được đo lường bởi tần suất chúng xuất hiện cùng nhau.

- Quan hệ cú pháp là một kiểu quan hệ đồng hiện đặc biệt, xuất hiện khi những vai trò thuộc về ngữ pháp của hai keyphrase trong câu xác định được. Nhãn, hướng, trọng số của cạnh trong trường hợp này có thể khác nhau tùy thuộc vào tri thức miền và kỹ thuật phân tích cú pháp câu.

Định nghĩa 9 *(Đồ thị keyphrase đơn giản biểu diễn câu truy vấn)*

Đánh trọng số cho các đỉnh của đồ thị

Mỗi đỉnh keyphrase k của đồ thị được gán một trọng số $w(k,d)$ để đánh giá mức độ hữu ích, tầm quan trọng của keyphrase trong việc phản ánh nội dung tài liệu d đó.

Tần số xuất hiện của keyphrase k **trong tài liệu** d, phản ánh mức độ quan trọng của keyphase trong tài liệu đang xét, ký hiệu là $tf(k,d)$, được tính bởi Công thức (4.1):

$$tf(k,d) = c + (1-c)\frac{n(k,d)}{\max(\{n(k',d)|k' \in d\})} \tag{4.1}$$

trong đó $n(k,d)$ là số lần keyphrase k xuất hiện trong tài liệu d, tham số $c \in [0,1]$ đóng vai trò làm giá trị tối thiểu cho trọng số tf của một keyphrase bất kỳ, việc này giúp giảm bớt sự chênh lệch trọng số tf giữa các keyphrase.

Độ chuyên biệt của keyphrase k **trong kho tài liệu** D, ký hiệu là $idf(k,D)$, được tính bởi Công thức (4.2).

$$\mathrm{idf}(k, D) = \log\left(\frac{|D|}{1 + |\{d \in D, k \in d\}|}\right) \tag{4.2}$$

trong đó $|D|$ là tổng số tài liệu trong kho và $|\{d \in D, k \in d\}|$ là số lượng tài liệu mà keyphrase k có xuất hiện.

Tầm quan trọng của keyphrase k **trong tài liệu** d **dựa vào vị trí xuất hiện của** k **trong tài liệu**, ký hiệu $\mathrm{ip}(k, d)$, được xác định bởi Công thức (4.3). Trọng số ip được dùng để đánh giá mức độ quan trọng của keyphrase dựa vào vị trí xuất hiện của keyphrase đó trong tài liệu.

$$\mathrm{ip}(k, d) = a + (1 - a)\frac{\sum_{i \in A} w_i}{\sum_i w_i} \tag{4.3}$$

trong đó,w_i là trọng số phản ánh độ quan trọng của phần nội dung thứ i trong cấu trúc tài liệu với ràng buộc $w_i \in [0, 1]$ và tham số $a = \max(w_i | i \in A)$ chính là trọng số của phần nội dung quan trọng nhất mà k xuất hiện và đây cũng sẽ là giá trị tối thiểu cho $\mathrm{ip}(k, d)$. Ta gọi tập hợp chứa chỉ số của tất cả phần nội dung mà k xuất hiện trong d là $A = \{x | n_x(k, d) > 0\}$ và $n_x(k, d)$ là số lần xuất hiện của keyphrase k trong phần nội dung x. Số phần nội dung (vị trí) và trọng số của mỗi phần có thể khác nhau tùy vào mỗi loại tài liệu.

Trọng số được gán cho đỉnh keyphrase k của đồ thị $\mathrm{docKG}(d)$, biểu thị giá trị ước tính về tính hữu ích của keyphrase k trong việc mô tả tài liệu d và được tính theo Công thức (4.4) như sau:

$$w(k, d) = \mathrm{tf}(k, d) \times \mathrm{idf}(k, D) \times \mathrm{ip}(k, d) \tag{4.4}$$

4.3 Đánh giá độ tương đồng ngữ nghĩa giữa tài liệu và câu truy vấn

Đánh giá độ liên quan giữa câu truy vấn và tài liệu được thực hiện bằng cách tính toán độ tương đồng ngữ nghĩa giữa hai đồ thị keyphrase biểu diễn chúng. Đồ thị keyphrase bao gồm các keyphrase và quan hệ tạo thành, nên phương hướng để thực hiện việc đo độ giống nhau về ngữ nghĩa giữa hai đồ thị là tìm ra độ tương đồng giữa các đỉnh keyphrase và giữa các cạnh quan hệ có trong hai đồ thị đó.

Định nghĩa 10 *(Khả năng đạt được)*

Định nghĩa 11 *(Độ tương đồng tiên đề)*

Khi k' có khả năng đạt được trực tiếp từ k bởi quan hệ $r \in R_{KK}$, bộ ba (k, r, k') có thể được gán một số thực trong khoảng giá trị $[0.0...1.0]$, được ký hiệu như $\mathrm{val}(k, r, k')$. Giá trị này đại diện cho *độ tương đồng tiên đề (axiomatic similarity degree)* của k và k' theo r.

Định nghĩa 12 *(Trọng lượng của đường đi)*

Cho $O = (K, R_{KK})$ *là một mô hình con đạt được từ ontology miền CK-ONTO biểu diễn một miền tri thức đặc biệt* $\mathbb{K}$*, hai keyphrase* $k, k' \in K$ *và một đường đi p có chiều dài n,* $p = (k_0 r_{s_1} k_1, k_1 r_{s_2} k_2, ...k_{n-1} r_{s_n} k_n)$ *từ* $k = k_0$ *đến* $k' = k_n$ *trong* O*. Cho hàm* $val : K \times R_{KK} \times K \to [0, 1]$ *xác định độ tương đồng tiên đề của một cặp keyphrase trong* K *theo một quan hệ trong* R_{KK}*. Trọng lượng của đường đi* p *được định nghĩa bởi Công thức (4.5) :*

$$V(k_0 r_{s_1} k_1, k_1 r_{s_2} k_2, ...k_{n-1} r_{s_n}) = \prod_{0}^{n-1} val(k_i, r_{s_i}, k_{i+1}) \tag{4.5}$$

Định nghĩa 13 *(Độ tương đồng ngữ nghĩa giữa hai keyphrase)*

Với mọi $k, k' \in K$*, độ tương đồng ngữ nghĩa giữa hai keyphrase k, k' được xác định thông qua hàm* $\alpha : K \times K \to [0, 1]$ *và được định nghĩa như sau:*

- $\alpha(k, k') = 1$ nếu $k = k'$
- $\alpha(k, k') = 0$ nếu $k \bar{P} k'$ tức là k' không có khả năng đạt được (not reachable) từ k
- $\alpha(k, k') = Max(\{V(P) | P$ là một đường đi từ k đến k'$\})$

Định nghĩa 14 *(Đường đi có trọng lượng lớn nhất)*

Độ tương đồng ngữ nghĩa giữa hai quan hệ

Đặt $\beta : T_R \cup T_S \times T_R \cup T_S \to [0, 1]$ là một ánh xạ cho phép đánh giá mức độ giống nhau về nghĩa giữa hai quan hệ. T_R là tập hợp các tên quan hệ được tìm thấy trong R_{KK} và T_S là tập tên của những quan hệ cú pháp giữa các keyphrase. Mặc dù biểu diễn của hàm này có thể được xác định tùy ý (thậm chí các giá trị của beta có thể được chọn thủ công), một số ràng buộc nên được xem xét, ví dụ như:

- $\forall r \in T_R \cup T_S, \beta(r, r) = 1$
- $\forall r, r' \in T_R \cup T_S, \beta(r, r') = \beta(r, r')$
- $\beta(\text{synonymy, abbreviation}) = 1.$
- Các quan hệ trong cùng một nhóm (chẳng hạn như những quan hệ thuộc nhóm Hierarchical relation) nên có độ giống nhau về mặt ngữ nghĩa cao hơn so với các quan hệ trong các nhóm khác nhau.

Định nghĩa 15 *(Phép chiếu KG)*

Đặt $H = (V_H, E_H, \phi_H, l_{E_H})$ *và* $G = (V_G, E_G, \phi_G, l_{E_G})$ *là hai đồ thị keyphrase được định nghĩa trên* $O = (K, R_{KK})$ *của CK-ONTO. Một phép chiếu KG từ H đến G là một cặp có thứ tự* $\Pi = (f, g)$ *của hai ánh xạ* $f : E_H \to E_G$, $g : V_H \to V_G$ *thỏa những điều kiện sau:*

- f và g là những hàm đơn ánh

- Phép chiếu bảo toàn "quan hệ kề" giữa các đỉnh của H, nghĩa là với mọi $e \in E_H, g(adj_i(e)) = adj_i(f(e))$, $adj_i(e)$ cho biết đỉnh thứ i kề với cạnh e. Nếu hai đỉnh kề kề nhau trong H thì các đỉnh tương ứng của nó cũng kề nhau trong G.

- $\forall e \in E_H, \beta(l_{E_H}(e), l_{E_G}(f(e))) \neq 0$

- $\forall k \in V_H, \alpha(k, g(k)) \neq 0$

Định nghĩa 16 *(Phép chiếu bộ phận)*

Tồn tại một phép chiếu bộ phận từ đồ thị keyphrase H tới đồ thị keyphrase G nếu và chỉ nếu tồn tại một phép chiếu KG từ H', một đồ thị keyphrase con của H (H' $\leq$ H), tới G.

Công thức (4.6) bên dưới cho phép lượng giá một phép chiếu KG từ H đến G, trong đó H là đồ thị của câu truy vấn và G là đồ thị biểu diễn tài liệu.

Định nghĩa 17 *(Mô hình lượng giá)*

Đặt $H = (V_H, E_H, \phi_H, l_{E_H})$ *là một đồ thị keyphrase đơn giản biểu diễn cho câu truy vấn q,* $G = (V_G, E_G, \phi_G, l_{E_G}, w_{V_G}, w_{E_G})$ *là một đồ thị keyphrase có trọng số (hoặc đồ thị keyphrase có trọng số đầy đủ) biểu diễn cho tài liệu d và* $H' \leq H$. *Một mô hình lượng giá cho phép chiếu bộ phận* Π *từ H' đến G được định nghĩa như Công thức (4.6) sau (có tỉ lệ về khoảng [0,1]) :*

$$v(\Pi) = \frac{|V_{H'}|/|V_H| \sum_{k \in V_{H'}} w_{V_G}(g(k)) \cdot \alpha(k, g(k)) + \sum_{e \in E_{H'}} \beta(e, f(e)) \cdot w_{E_G}(e)}{|V_{H'}| + |E_{H'}|} \tag{4.6}$$

Định nghĩa 18 *(Độ tương đồng ngữ nghĩa giữa câu truy vấn và tài liệu)*

Đặt H là một đồ thị keyphrase của câu truy vấn q, G là một đồ thị keyphrase của tài liệu d. Độ tương đồng ngữ nghĩa giữa câu truy vấn q và tài liệu d, cũng chính là độ tương đồng ngữ nghĩa giữa hai đồ thị H và G biểu diễn chúng, là một giá trị thuộc khoảng [0,1], được định nghĩa như Công thức (4.7) sau:

$$Rel(q, d) = Max(\{v(\Pi) | \Pi \textit{là một phép chiếu bộ phận từ H' đến G}, H' \leq H)\} \tag{4.7}$$

4.4 Kết chương

Chương 4 đề xuất một phương pháp biểu diễn tài liệu với tiếp cận là dựa trên ontology và biểu diễn văn bản bằng đồ thị. Để khắc phục những hạn chế trong việc biểu diễn tài liệu từ những mô hình truyền thống, đề tài đã nỗ lực thay đổi cách biểu diễn cho tài liệu dựa trên tiếp cận mới, trong đó có xem xét đến việc kết hợp thông tin cấu trúc và thông tin ngữ nghĩa trong các văn bản, nhằm làm tăng hiệu quả trong

biểu diễn và tìm kiếm. Trên cở sở các phương pháp biểu diễn nêu trên, đề tài đề xuất một kỹ thuật đo lường độ tương quan ngữ nghĩa giữa tài liệu và câu truy vấn, trong đó xây dựng được một số thuật giải cùng với những xử lý cơ bản nhằm giải quyết các vấn đề chính được đặt ra như sau: 1) Tính khoảng cách ngữ nghĩa giữa các keyphrase dựa trên việc khai thác nguồn tri thức ontology miền dựng sẵn, 2) So khớp đồ thị keyphrase, trên cơ sở đó đo lường mức độ liên quan giữa tài liệu và câu truy vấn.

Chương 5 HỆ QUẢN LÝ CƠ SỞ TÀI LIỆU VĂN BẢN THEO NGỮ NGHĨA: MỘT GIẢI PHÁP THIẾT KẾ HỆ THỐNG VÀ CÁC ỨNG DỤNG

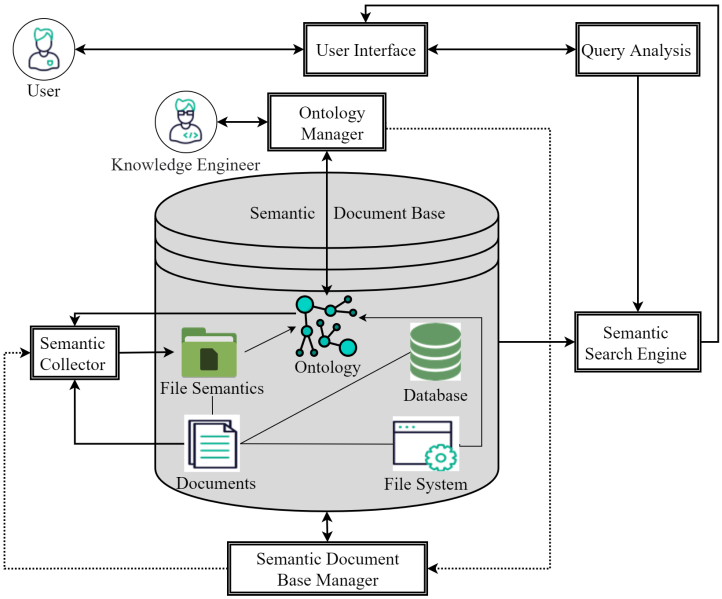
Chương 5 sẽ trình bày giải pháp thiết kế và xây dựng một lớp hệ thống ứng dụng gọi là "Hệ quản lý cơ sở tài liệu văn bản theo ngữ nghĩa", bằng cách đưa ra những đặc trưng cơ bản của hệ thống để phân biệt với những loại hệ thống khác, Kiến trúc hệ thống, Quy trình xây dựng và đặt ra một số vấn đề kỹ thuật cần phải giải quyết. Yêu cầu sử dụng của hệ thống bao gồm các tác vụ chính là tổ chức lưu trữ, quản lý và tìm kiếm chọn lọc, đặc biệt là khả năng tìm kiếm dựa trên tri thức của lĩnh vực hay theo ngữ nghĩa liên quan đến nội dung của tài liệu. Kết quả nghiên cứu đã được công bố trong các công trình [CT1], [CT2], [CT3] của tác giả.

5.1 Giới thiệu

5.2 Hệ quản lý cơ sở tài liệu văn bản theo ngữ nghĩa

Hệ quản lý cơ sở tài liệu văn bản theo ngữ nghĩa (Semantic Document Base System, viết tắt là SDBS) là một hệ thống máy tính tập trung vào việc sử dụng các kỹ thuật trí tuệ nhân tạo để tổ chức kho tài liệu văn bản trên máy tính một cách hiệu quả, trong đó cố gắng quản lý được các thông tin ngữ nghĩa liên quan đến nội dung của tài liệu cũng như hỗ trợ biểu diễn và xử lý ngữ nghĩa trong quá trình tìm kiếm thông tin (tài liệu) dựa trên tri thức miền ứng dụng. Hệ thống tích hợp nhiều thành phần, trong đó có một kho lưu trữ tài liệu (cơ sở tài liệu) thuộc về một miền tri thức nào đó, và việc lập chỉ mục cho kho tài liệu dựa trên nội dung được yêu cầu, cùng với một động cơ tìm kiếm theo ngữ nghĩa trong phạm vi tri thức của hệ thống.

Về mặt kiến trúc, các hệ thống SDBS có thể gồm các thành phần như Cơ sở tài liệu theo ngữ nghĩa (Semantic Document Base), Bộ lập chỉ mục (Semantic Collector and Indexing), Động cơ tìm kiếm theo ngữ nghĩa (Semantic Search engine), Mô-đun quản lý cơ sở tài liệu (Semantic Doc Base Manager) bao gồm cả Mô-đun quản lý tri thức (Ontology Manager), Giao diện người dùng (User Interface), Bộ phân tích câu truy vấn của người dùng (Query Analyzer). Hệ thống phục vụ cho hai loại người dùng chủ yếu là người dùng phổ thông (user) và người kỹ sư tri thức (knowledge engineer). Hình 5.2 là sơ đồ kiến trúc thông dụng của một Hệ quản lý cơ sở tài liệu văn bản theo ngữ nghĩa.



Hình 5.2 Kiến trúc của một hệ thống SDBS

5.3 Hệ thống quản lý kho tài nguyên học tập về lĩnh vực Khoa học máy tính

Mục tiêu của ứng dụng là xây dựng một hệ thống quản lý kho tài nguyên học tập lĩnh vực Khoa học máy tính trong phạm vi một trường đại học. Ứng dụng đáp ứng các yêu cầu sau đây: *Về yêu cầu tìm kiếm, chọn lọc tài liệu*, hệ thống hỗ trợ tìm kiếm cơ bản theo từ khóa và tìm kiếm theo ngữ nghĩa.*Về yêu cầu quản lý cơ bản*, hệ thống cho phép quản lý quy trình nghiệp vụ xử lý và thao tác với tài liệu, hệ thống thư mục quy chuẩn, cơ sở dữ liệu của kho tài liệu, các biểu diễn ngữ nghĩa dưới dạng đồ thị keyphrase, đặc biệt là việc xây dựng và tổ chức quản lý lưu trữ, cập nhật, tìm kiếm trên ontology CK-ONTO, cũng như mối liên hệ giữa ontology với các thành phần khác.

Kết quả trong Bảng 5.2 cho thấy phương pháp đề xuất có kết quả tìm kiếm văn bản vượt trội hơn hẳn so với các chương trình dùng mô hình biểu diễn văn bản truyền thống như mô hình không gian vector, mô hình xác suất hay các mô hình ngôn ngữ.

Có thể thấy độ chính xác của các phương pháp SDB khá cao trong số 20 kết quả đầu nhưng giảm nhanh chóng và hội tụ về mức trung bình trong thử nghiệm trước khi xét thêm các tài liệu được xếp hạng thấp trong kết quả thử nghiệm. Ngược lại với Lucene, độ chính xác ở các ngưỡng giới hạn khác nhau không có quá nhiều khác biệt.

Bảng 5.2 Hiệu quả tìm kiếm của Hệ thống quản lý kho tài nguyên học tập về lĩnh vực Khoa học máy tính trên kho thử nghiệm gồm 1000 tài liệu và 100 câu truy vấn (theo phần trăm)

Mô hình	**Độ chính xác**	**Độ bao phủ**	**Độ F**
SDB + fulldocKG	**75.1**	**79.3**	**77.1**
SDB + docKG	**71.3**	**70.9**	**71.1**
VSM(Lucene)	47.2	**88.7**	61.6
MySQL FTS	15.7	68.5	25.5
BM25	44.1	71.5	54.6
LDA	37.6	65.8	47.8

Bảng 5.3 Hiệu quả tìm kiếm của Hệ thống quản lý kho tài nguyên học tập về lĩnh vực Khoa học máy tính trên kho thử nghiệm gồm 10.000 tài liệu và 100 câu truy vấn (theo phần trăm)

Mô hình	**P@20**	**P@100**	**P@1k**
SDB + fulldocKG	**82.1**	**75.3**	**70.4**
SDB + docKG	**75.3**	**73.9**	**68.1**
VSM(Lucene)	45.5	44.7	47.1
BM25	44.0	43.5	41.3
LDA	43.4	41.1	37.5

5.4 Hệ thống tìm kiếm tin bài tuyển dụng ngành Công nghệ Thông tin

Hệ thống tìm kiếm tin bài tuyển dụng đặt mục tiêu nhằm hỗ trợ cho những người dùng đang có nhu cầu tìm việc làm trong ngành công nghệ thông tin. Hệ thống giúp người dùng có thể tìm kiếm những tin bài tuyển dụng có mô tả công việc phù hợp nhất trên một số trang tin bài tuyển dụng. Tổng cộng có 100 mẫu truy vấn sẽ được đưa vào thử nghiệm và đánh giá. Có tổng cộng khoảng hơn 2500 tin bài tuyển dụng được thu thập ở dạng HTML, các tin bài này sau đó được chuyển sang dạng chữ trơn (plain text) trước khi đưa vào hệ thống xử lý.

5.5 Hệ thống tìm kiếm và chọn lọc tin bài trên các báo điện tử

Đề tài đã xây dựng phần mềm Web có khả năng thu thập tin bài của các trang báo điện tử, quản lý kho tin bài thu thập được, hỗ trợ tìm kiếm, chọn lọc các tin bài về tỉnh Bình Dương cho các lĩnh vực Lao động – Việc làm, Đầu tư công – Đầu tư nước ngoài và thực hiện điểm tin, tổng hợp tin bài theo nhiều tiêu chí khác nhau. Ứng dụng phục vụ cho nhu cầu thực tế của Phòng Báo chí và Xuất bản của Sở Thông tin và Truyền thông Bình Dương. Ứng dụng đáp ứng các yêu cầu sau đây: *Về yêu cầu quản lý cơ bản*, hệ thống cho phép quản lý cấu trúc thông tin của các trang báo điện tử, quản lý Cơ

Bảng 5.5 Hiệu quả tìm kiếm của Hệ thống tìm kiếm tin bài tuyển dụng ngành Công nghệ Thông tin (theo phần trăm)

Mô hình	**Độ chính xác**	**Độ bao phủ**	**Độ F**
SDB + fulldocKG	**77.1**	**77.8**	**77.4**
SDB + docKG	**70.3**	**71.9**	**71.1**
VSM(Lucene)	8.7	**98.5**	16.0
VSM(Lucene) + CKTokenizer	43.7	58.5	50.0
VSM(Lucene)+ CKQe	45.1	70.3	54.9
BM25	14.1	82.5	24.1
LDA	33.7	60.8	46.1

sở dữ liệu tin bài, quản lý phân loại tin bài theo lĩnh vực; Đối với *nhóm chức năng thu thập các tin bài*, đề tài xây dựng được một chương trình hỗ trợ quản lý và lập lịch cho quá trình thu thập tin tức tự động. Tin bài thu thập có thể được phân loại tự động vào các lĩnh vực theo cấu trúc của trang báo cũng như lĩnh vực phân loại tổng quát được tổ chức theo ý của người dùng; Đặc biệt nhất là *Phân hệ ứng dụng với giao diện web hỗ trợ tìm kiếm, chọn lọc thông tin về tỉnh Bình Dương trên các báo mạng cho lĩnh vực Lao động – Việc làm và Đầu tư công – Đầu tư nước ngoài, tổng hợp và trích xuất báo cáo điểm tin hàng ngày.*

Kết quả thử nghiệm cho chức năng tìm kiếm theo ngữ nghĩa với 50 câu truy vấn liên quan đến Lao động – Việc làm và Đầu tư công – Đầu tư nước ngoài, trên tập dữ liệu gồm 1000 các tin bài thu thập được thuộc nhiều lĩnh vực khác nhau, ngưỡng chặn là 0.5 cho thấy độ chính xác trung bình của hệ thống là 83.1%, độ bao phủ trung bình đạt 82.1%. Hệ thống tìm được hầu hết các tin bài liên quan đến nội dung tìm kiếm và sắp xếp theo mức độ liên quan giảm dần. Tốc độ xử lý tìm kiếm trong thời gian chấp nhận được. Chức năng tìm kiếm theo từ khóa (sử dụng mô hình Không gian vector truyền thống của Lucence) trên Bộ dữ liệu thử nghiệm tương tự cho kết quả với độ chính xác trung bình là 60%, độ phủ trung bình là 78% và đặc biệt là có tốc độ xử lý nhanh, vượt trội hơn hẳn chức năng còn lại.

5.6 Kết chương

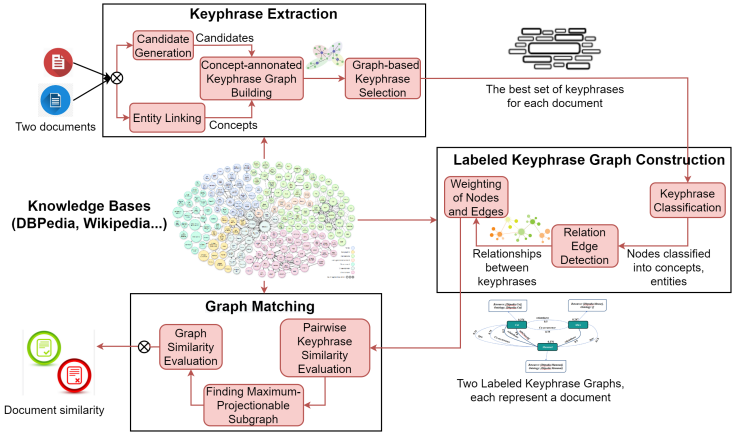
Luận án đã nêu lên các ưu thế và lợi ích của việc nghiên cứu phát triển các mô hình cùng với các thuật giải tự động dựa trên tri thức thông qua việc thiết kế, cài đặt và xây dựng được 03 ứng dụng thử nghiệm. Luận án đã tiến hành thực nghiệm trên các Bộ dữ liệu mẫu để so sánh và đánh giá hiệu quả tìm kiếm của hệ thống. Các kết quả thực nghiệm bước đầu cho thấy giải pháp đã đề xuất là khả quan và có khả năng ứng dụng tốt.

Chương 6 ĐO LƯỜNG MỨC ĐỘ TƯƠNG ĐỒNG NGỮ NGHĨA GIỮA HAI TÀI LIỆU VỚI TRI THỨC TỔNG QUÁT DỰA TRÊN ĐỒ THỊ KEYPHRASE

Những đề xuất về một phương pháp mới giải quyết bài toán đo lường độ tương đồng ngữ nghĩa giữa hai tài liệu, được trình bày trong Chương 6 của luận án, và đã được công bố trong các công trình [CT4][CT5] của tác giả.

6.1 Giới thiệu

Hình 6.1 cung cấp một cách nhìn tổng quan về cách tiếp cận của đề tài cho bài toán này.



Hình 6.1 Quy trình tính toán độ tương đồng giữa hai tài liệu

6.2 Mô hình hóa nội dung tài liệu bằng đồ thi dựa trên tri thức

Định nghĩa 19 *(Bộ từ vựng DBpedia)*

Một bộ từ vựng DBpedia (DBpedia vocabulary) hoặc gọi đơn giản là bộ từ vựng (vocabulary), là một hệ thống gồm có ba thành phần (T_O, T_R, R_{CC}) *trong đó:*

- T_O và T_R là hai tập hợp hữu hạn, rời nhau đôi một.

- $T_O = T_C \cup T_E$ là một tập hợp bao gồm đầy đủ các khái niệm (cũng được gọi là lớp) và tất cả các thực thể có trong DBpedia (DBpedia's concepts, entities).

- T_R là tập hợp các tên quan hệ (hoặc gọi là các ký hiệu quan hệ) được tìm thấy trên DBpedia.

- $R_{CC} \subseteq T_O \times T_R \times T_O$ là một tập hợp các quan hệ ngữ nghĩa giữa những khái niệm và thực thể được tìm thấy trên DBpedia.

Định nghĩa 20 *(Đồ thị keyphrase có gán nhãn)*

Cho một tài liệu d , một đồ thị keyphrase có gán nhãn (labeled keyphrase graph), biểu diễn cho tài liệu d, ký hiệu là labdocKG(d)*, được định nghĩa trên Bộ từ vựng DBPedia* $O = (T_O, T_R, R_{CC})$*, là một bộ* $G = (V, E, \phi, l_V, l_E, w_V, w_E)$ *thỏa các điều kiện sau:*

- (V, E, ϕ) là một đa đồ thị, hữu hạn, có hướng được gọi là đồ thị cơ sở của G, ký hiệu là graph(G). V là một tập hữu hạn, khác rỗng, bao gồm các keyphrase được đề cập đến trong tài liệu, cũng là tập các đỉnh của đồ thị. E là tập các cạnh quan hệ của đồ thị. $\phi : E \to \{(x, y)|(x, y) \in V^2, x \neq y\}$ là một hàm ánh xạ mỗi cạnh đến một cặp có thứ tự của hai đỉnh khác nhau. Mỗi cạnh biểu thị cho một mối quan hệ ngữ nghĩa hoặc quan hệ cú pháp giữa hai đỉnh kề với nó.

Hai đỉnh $k1, k2 \in V$ được kết nối với nhau nếu có tồn tại một quan hệ $r \in T_R$ sao cho $(k1, r, k2) \in R_{CC}$. Ngoài những mối quan hệ ngữ nghĩa, hai đỉnh keyphrase cũng có thể được kết nối với nhau nếu có tồn tại một hình thức nào đó thể hiện mối quan hệ cú pháp giữa chúng chẳng hạn như quan hệ đồng xuất hiện hoặc quan hệ ngữ pháp trong câu.

- $l_V : V \to \wp(T_O)$ and $l_E : E \to T_R \cup T_S$ là hai hàm gán nhãn cho những đỉnh và cạnh của graph(G). Mỗi đỉnh có thể được gán nhiều nhãn nên được gọi là multi-label node. Mỗi cạnh $e \in E$ cũng được gán nhãn bởi một tên quan hệ $l_E(e) \in T_R \cup T_S$. T_S là tập tên của những quan hệ về cú pháp được dùng để gán nhãn cho các cạnh này. $\wp(T_O)$ là tập hợp tất cả các tập con của T_O.

- $w_V : V \to [0, 1]$ and $w_E : E \to [0, 1]$ là hai ánh xạ được sử dụng cho việc đánh trong số cho các đỉnh và các cạnh của graph(G). Các trọng số phản ánh mức độ quan trọng của những đối tượng này trong việc đặc tả nội dung nồng cốt của văn bản.

Định nghĩa 21 *(Đồ thị keyphrase con có gán nhãn)*

Quy trình xây dựng đồ thị keyphrase biểu diễn một tài liệu bao gồm ba bước chính như: tạo ứng viên (candidate generation), liên kết thực thể (entity linking) và lựa chọn tự động dựa trên đồ thị (graph-based automatic selection). Trong các nghiên cứu về rút trích keyphrase trước đây, hai bước đầu tiên đã trở thành những bước thường được sử dụng để giải quyết bài toán này.

Cho một tài liệu d cùng với một danh sách các ứng viên keyphrase K được rút trích từ bước "Tạo ứng viên"và một danh sách các khái niệm C được phát hiện bởi tác vụ

”Liên kết thực thể”, đề tài sẽ tạo ra một đồ thị keyphrase có chú thích khái niệm (đặt tên là concept-annotated keyphrase graph) biểu diễn cho tài liệu d. Đồ thị này được sử dụng để chọn lọc các keyphrase từ tập ứng viên. Những ứng viên được đánh giá là giàu tiềm năng khi chúng có nhiều liên kết với những khái niệm quan trọng trong tài liệu.

Định nghĩa 22 *(Đồ thị keyphrase có chú thích khái niệm)*

Một đồ thị keyphrase có chú thích khái niệm (concept-annotated keyphrase graph) biểu diễn cho tài liệu d, ký hiệu $\mathrm{aKG(d)}$*, là một bộ* $\mathrm{G} = (\mathrm{V_k}, \mathrm{V_c}, \mathrm{E}, \mathrm{w})$ *thỏa mãn các điều kiện sau:*

- $(\mathrm{V_k}, \mathrm{V_c}, \mathrm{E})$ là một đồ thị lưỡng phân, hữu hạn và vô hướng.

- $\mathrm{V_k} \subseteq \mathrm{K}$ là một tập không rỗng của những đỉnh ứng viên (candidate node), $\mathrm{V_c} \subseteq \mathrm{C}$ là một tập những đỉnh khái niệm (concept node). Tập đỉnh của đồ thị bao gồm: $\mathrm{V} = \mathrm{V_k} \cup \mathrm{V_c}, \mathrm{V_k} \cap \mathrm{V_c} \neq \emptyset$.

- E là tập hợp những cạnh vô hướng của đồ thị. Tập đỉnh của đồ thị được phân hoạch thành hai tập không rỗng, rời nhau $\mathrm{V_k}$ và $\mathrm{V_c}$, tương ứng với hai loại đỉnh khác nhau. Mỗi cạnh của đồ thị chỉ nối chính xác một đỉnh trong $\mathrm{V_k}$ với một đỉnh trong $\mathrm{V_c}$, tức là chỉ có thể đi từ một đỉnh ứng viên đến một đỉnh khái niệm.

- Đối với mỗi đỉnh ứng viên $\mathrm{k} \in \mathrm{V_k}$ và mỗi đỉnh khái niệm $\mathrm{c} \in \mathrm{V_c}$, có thể có một cạnh nối giữa k và c nếu và chỉ nếu k ”chứa”tên của khái niệm (concept name) hoặc nếu k chứa đề cập của c.

Nói cách khác, tên khái niệm hoặc đề cập của c là một chuỗi con được tìm thấy trong k. Đề cập (mention) của khái niệm c là một từ hoặc một cụm từ xuất hiện trong tài liệu được chú thích đến c bởi tác vụ Liên kết thực thể.

- $\mathrm{w} : \mathrm{V_k} \cup \mathrm{V_c} \rightarrow \mathbb{R}^+$ là một hàm đánh trọng số cho các đỉnh của đồ thị. Những trọng số này biểu diễn sự ước lượng về mức độ hữu ích của những ứng viên và khái niệm trong việc mô tả nội dung hay chủ đề mà tài liệu đang đề cập tới, nhằm phân biệt tài liệu d với những tài liệu khác trong tập văn bản.

Trọng số liên kết với đỉnh khái niệm $\mathrm{c} \in \mathrm{V_c}$ của đồ thị phản ánh tầm quan trọng của khái niệm đối với tài liệu đã cho. Trọng số này được đánh giá dựa trên tần số xuất hiện của tất cả các đề cập của c trong tổng thể tài liệu. Vì c là một khái niệm được chú thích từ một cơ sở tri thức nguồn (như Wikipedia chẳng hạn), do đó có nhiều trường hợp tên của khái niệm không xuất hiện đầy đủ trong tài liệu. Vì vậy, trọng số được tính toán thông qua các đề cập của khái niệm này, tức là bằng cách tính tổng số lần xuất hiện trong d của tất cả các đề cập được chú thích đến cùng một khái niệm là c.

Để chọn được một danh sách các keyphrase có sắp hạng theo mức độ ưu tiên, đề tài áp dụng thuật giải được trình bày trong [111] và có một số điều chỉnh như trong Thuật toán 4.

Cho trước một tập hợp các keyphrase được phát hiện ở bước 1, việc xây dựng một đồ thị keyphrase có gán nhãn của tài liệu được thực hiện theo 3 bước chính sau: phân loại keyphrase, phát hiện cạnh quan hệ và đánh trọng số cho các thành phần của đồ thị.

6.3 Đánh gía độ tương đồng giữa hai tài liệu dựa trên đồ thị

Định nghĩa 23 *(Độ tương đồng ngữ nghĩa giữa hai khái niệm)*

Độ tương đồng ngữ nghĩa giữa hai khái niệm (Pairwise concept similarity) c_i *và* c_j *được định nghĩa dựa trên Bộ từ vựng DBpedia* (T_O, T_R, R_{CC}) *theo Công thức (6.1) [53] sau:*

$$sim_{wpath}(c_i, c_j) = \frac{1}{1 + length(c_i, c_j) * k^{IC(c_{lcs})}} \tag{6.1}$$

trong đó, $length(c_i, c_j)$ cho biết chiều dài đường đi ngắn nhất giữa hai khái niệm c_i và c_j trong đồ thị O. Đường đi giữa hai khái niệm càng ngắn thì chúng được xem là càng giống nhau về mặt ngữ nghĩa.

Hàm lượng IC của một khái niệm c, ký hiệu là $IC(c)$, đo lường mức độ cung cấp thông tin của khái niệm và có thể được định nghĩa như sau: $IC(c) = -\log Prob(c)$ trong đó $Prob(c) = \frac{|entities(c)|}{N}$, N là tổng số thực thể có trong O và $entities(c)$ là tập hợp các thực thể có kiểu (type) là c.

Cũng trong Công thức (6.1), ký hiệu c_{lcs} đại diện cho tổ tiên chung gần nhất (Least Common Subsumer -LCS) của hai khái niệm c_i và c_j. Phần tử này được tìm thấy trong một cây phân cấp các khái niệm (taxonomy of concepts) của DBpedia vốn được xây dựng từ O nhờ vào quan hệ is-a. Tham số $k \in (0, 1]$ biểu thị cho tỷ trọng đóng góp của hàm lượng IC của đối tượng c_{lcs} vào độ tương đồng giữa hai khái niệm. Giá trị của tham số k được chọn thông qua thực nghiệm là 0.8.

Định nghĩa 24 *(Độ tương đồng ngữ nghĩa giữa hai thực thể)*

Đặt $In(e_i)$ *và* $In(e_j)$ *lần lượt là tập các "liên kết đến"(incoming link) trỏ đến trang Wikipedia của thực thể* e_i *và* e_j*. Độ tương đồng ngữ nghĩa giữa hai thực thể (Pairwise entity similarity)* e_i *và* e_j *được đánh giá qua Công thức (6.2)[35] như sau:*

$$rel(e_i, e_j) = 1 - \frac{\log(\max(|In(e_i)|, |In(e_j)|)) - \log(|In(e_i) \cap In(e_j)|)}{\log(|W|) - \log(\min(|In(e_i)|, |In(e_j)|))} \tag{6.2}$$

với W là tập các thực thể có trong cơ sở tri thức Wikipedia.

Định nghĩa 25 *(Độ tương đồng về mặt khái niệm giữa hai keyphrase)*

Cho hai keyphrase k_1 *và* k_2*, đặt* $C_1 = \{c_{11}, c_{12}, ..., c_{1p}\}$ *và* $C_2 = \{c_{21}, c_{22}, ..., c_{2q}\}$ *lần lượt là tập hợp các khái niệm có liên kết tới* k_1 *và* k_2*. Độ tương đồng "về mặt khái*

niệm"giữa hai keyphrase k_1 *và* k_2 *được xác định thông qua mức độ giống nhau của hai nhóm khái niệm (Groupwise concept similarity) liên kết với chúng theo Công thức (6.3) sau:*

$$gcs(k_1, k_2) = \max\{sim_{wpath}(c_{1i}, c_{2j}) | c_{1i} \in C_1, c_{2j} \in C_2\} \tag{6.3}$$

Định nghĩa 26 *(Độ tương đồng về mặt thực thể giữa hai keyphrase)*

Cho hai keyphrase k_1 *và* k_2*, đặt* $E_1 = \{e_{11}, e_{12}, ..., e_{1p}\}$ *và* $E_2 = \{e_{21}, e_{22}, ..., e_{2q}\}$ *lần lượt là tập hợp các thực thể có liên kết tới* k_1 *và* k_2*. Độ tương đồng "về mặt thực thể"giữa hai keyphrase* k_1 *và* k_2 *được xác định thông qua mức độ giống nhau của hai nhóm thực thể (Groupwise entity similarity) liên kết với chúng theo Công thức (6.4) sau:*

$$gec(k_1, k_2) = \max\{rel(e_{1i}, e_{2j}) | e_{1i} \in E_1, e_{2j} \in E_2\} \tag{6.4}$$

Cuối cùng, độ tương đồng ngữ nghĩa giữa hai keyphrase chính là giá trị lớn nhất trong hai độ tương đồng theo nhóm kể trên.

Định nghĩa 27 *(Độ tương đồng ngữ nghĩa giữa hai keyphrase) Cho hai keyphrase* k_1 *và* k_2*. Độ tương đồng ngữ nghĩa giữa hai keyphrase (Pairwise keyphrase similarity)* k_1 *và* k_2 *được xác định bởi Công thức (6.5):*

$$sim(k_1, k_2) = \max\{gcs(k_1, k_2), gec(k_1, k_2)\} \tag{6.5}$$

Việc tính toán độ tương đồng ngữ nghĩa giữa hai keyphrase có thể được thực hiện theo Thuật toán 5.

Định nghĩa 28 *(Độ tương đồng giữa hai quan hệ)*

Độ tương đồng giữa hai cạnh quan hệ (Relation Similarity) trong đồ thị keyphrase được định nghĩa đơn giản là một tỷ lệ giữa trọng số của cạnh có giá trị thấp hơn và trọng số cạnh có giá trị cao hơn.

Định nghĩa này cho phép đánh giá mức độ giống nhau giữa hai cạnh bất kỳ nào đó, bất kể loại quan hệ của hai cạnh là khác biệt nhau.

Một số định nghĩa cần thiết dùng cho việc so khớp đồ thị cũng như tính toán độ tương đồng giữa hai đồ thị được cho như sau:

Định nghĩa 29 *(Phép chiếu Labeled-KG)*

Cho $G = (V_G, E_G, \phi_G, l_{V_G}, l_{E_G}, w_{V_G}, w_{E_G})$ *và* $H = (V_H, E_H, \phi_H, l_{V_H}, l_{E_H}, w_{V_H}, w_{E_H})$ *là hai đồ thị keyphrase có gán nhãn. Một* phép chiếu Labeled-KG (Labeled-KG projection) *từ G đến H là một cặp có thứ tự* $\Pi = (f, h)$ *của hai ánh xạ* $f : V_G \to V_H$, $h : E_G \to E_H$*, thỏa những điều kiện sau:*

- f và h là những hàm đơn ánh.

- Phép chiếu bảo toàn "quan hệ kề" giữa các đỉnh của G , nghĩa là với mọi $e \in E_G, f(adj_i(e)) = adj_i(h(e))$, $adj_i(e)$ cho biết đỉnh thứ i kề với cạnh e.

- $\forall k \in V_G, sim(k, f(k)) \neq 0$, với $sim(k, f(k)) \in [0, 1]$ là độ tương đồng ngữ nghĩa giữa hai keyphrase k và f(k) theo Định nghĩa 27.

Định nghĩa 30 *(Phép chiếu bộ phận Labeled-KG) Cho G và H là hai đồ thị keyphrase có gán nhãn, tồn tại một* phép chiếu bộ phận Labeled-KG (Labeled-KG partial projection) *từ G đến H nếu và chỉ nếu tồn tại một phép chiếu Labeled-KG từ G', một sub labeled keyphrase graph (sublabKG) của G (G' $\leq$ G), tới H.*

Định nghĩa 31 *(Đồ thị con có thể chiếu được)*

Cho G và H là hai đồ thị keyphrase có gán nhãn. Một đồ thị keyphrase có gán nhãn g được gọi là đồ thị con có thể chiếu được (projectionable subgraph) của G và H nếu thỏa các điều kiện sau:

- $g \leq G$
- Tồn tại một phép chiếu Labeled-KG $\Pi = (f, h)$ từ g đến H.

Định nghĩa 32 *(Đồ thị con lớn nhất có thể chiếu được)*

Cho G và H là hai đồ thị keyphrase có gán nhãn. Một đồ thị keyphrase có gán nhãn g được gọi là đồ thị con lớn nhất có thể chiếu được (maximum-projectionable subgraph) của G và H, ký hiệu là $mps(G, H)$*, nếu thỏa các điều kiện sau:*

- g là một đồ thị con có thể chiếu được của G và H.
- Không có một đồ thị con có thể chiếu được g' nào khác của G và H mà $|V_{g'}| > |V_g|$.

Định nghĩa 33 *(Độ tương đồng ngữ nghĩa giữa hai đồ thị)*

Cho $G = (V_G, E_G, \phi_G, l_{V_G}, l_{E_G}, w_{V_G}, w_{E_G})$ *và* $H = (V_H, E_H, \phi_H, l_{V_H}, l_{E_H}, w_{V_H}, w_{E_H})$ *là hai đồ thị keyphrase có gán nhãn. Độ tương đồng ngữ nghĩa giữa hai đồ thị (similarity between two graphs) G và H được định nghĩa bởi Công thức (6.6) như sau:*

$$Sim(G, H) = \beta \times \frac{\sum_{k}^{V_g} sim(k, f(k)) \times w_V(k, f(k))}{\max(|V_G|, |V_H|)} + (1 - \beta) \times \frac{\sum_{e}^{E_g} w_E(e, h(e))}{\max(|E_G|, |E_H|} \quad (6.6)$$

trong đó, $w_V(k, f(k)) = \frac{\min(w_{V_G}(k), w_{V_H}(f(k))}{\max(w_{V_G}(k), w_{V_H}(f(k))}$,

$w_E(e, h(e)) = \frac{\min(w_{E_G}(e), w_{E_H}(h(e)))}{\max(w_{E_G}(e), w_{E_H}(h(e)))}$,

Ký hiệu $g = mps(G, H)$ đại diện cho đồ thị con lớn nhất có thể chiếu được của G và H. $\beta \in (0, 1)$ là một hệ số do người dùng xác định, cho biết tỷ lệ phần trăm đóng góp của từng thành phần bao gồm hai giá trị là "so khớp đỉnh"và "so khớp cạnh"vào kết quả so khớp đồ thị cuối cùng. Có thể tồn tại nhiều phép chiếu Labeled-KG từ g đến H. Phép chiếu được chọn là phép chiếu có $\sum_k^{V_g} sim(k, f(k)) \times w_V(k, f(k))$ đạt giá trị lớn nhất.

Quá trình trích xuất một "đồ thị con lớn nhất có thể chiếu được"từ hai đồ thị là một nhiệm vụ khó khăn và phức tạp, đòi hỏi phải xem xét riêng biệt. Một phương pháp tiếp cận heuristic được đề xuất cho việc so khớp và ước tính độ tương đồng giữa hai đồ, như được trình bày trong Thuật toán 6.

6.4 Thực nghiệm đánh giá kỹ thuật đo lường độ tương đồng ngữ nghĩa giữa hai tài liệu

Các thực nghiệm mang tính toàn diện được yêu cầu, không chỉ để đánh giá tính hiệu quả của phương pháp đề xuất mà còn để xác định các tham số thích hợp tham gia vào các công thức và thuật toán cơ sở. Kết quả tốt nhất của phương pháp đề xuất, đạt được bằng cách sử dụng cấu hình trong Bảng 6.3, sẽ được so sánh với các phương pháp hiện đại khác được nghiên cứu cho bài toán đo lường mức độ tương đồng ngữ nghĩa giữa hai tài liệu.

Bảng 6.3 Các tham số và giá trị tối ưu của chúng

Tham số	**Giá trị**
Mẫu ứng viên (Candidate pattern)	< JJ|JJR|JJS > ∗ < NN|NNS|NNP|NNPS > +
Công cụ liên kết thực thể (Entity Linker)	TagMe with confident score 0.11
Tỉ lệ giảm trọng số α	1.5
Trọng số đỉnh keyphrase	closeness centrality+tf+idf*pagerank (cho kết quả tốt nhất so với các cách tổ hợp khác)
Loại cạnh	Co-occurr, relatedness_entity, subClass, entity_type_concept, alt_name
Hệ số β trong Công thức (6.6)	0.71

Bảng 6.4 trình bày kết quả của các phương pháp đã thử nghiệm và được sắp xếp theo điểm số Pearson tương ứng của chúng.

Kết quả thử nghiệm cho thấy rằng cách tiếp cận của đề tài vượt trội (về điểm số Pearson) hơn các phương pháp dựa trên mô hình truyền thống, kể cả RoBERTa, MPNet

Bảng 6.4 Kết quả thực nghiệm trên Bộ dữ liệu LP50

Phương pháp	**Độ đo Pearson r**
CSA	0.62
GED	0.63
MPNet	0.63
CSA + LSA	0.65
Roberta	0.65
ESA paper	0.72
CSA + ESA	0.72
ConceptGraphSim	0.74
WikiWalk + ESA	0.77
ConceptGraphSim + ESA	0.78
KGM (+TagMe)	**0.79**
KGM (+LE-TagMe)	**0.80**
Concepts Learned	0.80
KGM (+LE-TagMe) + ESA reimplement	**0.81**

và có thể đạt được hiệu suất ngang bằng với các phương pháp state-of-the-art như Concepts Learned, ConceptGraphSim và ESA. Hơn nữa, khi kết hợp kết quả đánh giá độ tương đồng đạt được từ phương pháp nguyên mẫu cùng với kết quả đánh giá có được thông qua việc cài đặt lại ESA (ký hiệu KGM+ESA), đề tài thậm chí có thể vượt lên trên Concepts Learned một phần trăm nhỏ.

Khi xem xét kỹ hơn một số trường hợp mà hệ thống trả về kết quả đánh gía độ tương tự giữa hai tài liệu khác biệt nhiều so với đánh giá của con người, đề tài nhận thấy rằng: với cách tiếp cận dựa trên đồ thị keyphrase, giải pháp chưa cho kết quả đo lường tốt trên các cặp tài liệu có cùng chủ đề nhưng khác nhau về thời gian và không gian diễn ra sự kiện được đề cập trong các tài liệu.

6.5 Kết chương

Trong Chương 6, đề tài đã giới thiệu một số mô hình biểu diễn tài liệu dựa trên đồ thị cùng với một chiến lược ”giàu tri thức”để tạo ra các cấu trúc giải nghĩa cho nội dung văn bản. Đỉnh keyphrase sẽ được liên kết tới các khái niệm hoặc thực thể trong DBpedia và Wikipedia nhằm tạo điều kiện cho việc hiểu rõ hơn về ý nghĩa diễn đạt của từ vựng trong văn bản. Ngoài ra, trong nghiên cứu này, một kỹ thuật so khớp đồ thị mới được đề xuất nhằm đánh giá khoảng cách ngữ nghĩa giữa hai tài liệu. Kỹ thuật này cũng có thể được áp dụng cho nhiều tác vụ liên quan đến văn bản khác bao gồm truy xuất tài liệu, phân loại tài liệu và xếp hạng thực thể.

KẾT LUẬN

KẾT QUẢ ĐẠT ĐƯỢC

Về mặt khoa học

Kết quả thứ nhất là đề xuất một phương pháp mới cho việc giải quyết bài toán Tìm kiếm theo ngữ nghĩa trên một kho tài liệu văn bản có liên quan đến một miền tri thức nhất định nào đó.

Đề tài nghiên cứu phương pháp tìm kiếm tài liệu theo hướng cải tiến độ chính xác và độ bao phủ, không đặt vấn đề về hiệu năng (thời gian xử lý truy vấn, kích thước chỉ mục, xử lý phân tán) của hệ thống khi được triển khai thực tế.

Đề tài đã nỗ lực cải thiện hiệu quả tìm kiếm thông qua việc nghiên cứu các phương pháp biểu diễn ngữ nghĩa cho tài liệu cùng với kỹ thuật tính toán độ tương đồng ngữ nghĩa giữa tài liệu và câu truy vấn. Giải pháp được đề xuất đi theo tiếp cận dựa trên đồ thị và tận dụng một ontology miền với độ mịn cao, được kiểm soát tốt để làm cơ sở cải thiện hiệu quả tìm kiếm các tài liệu thuộc miền. Phương pháp mới được đề xuất bao gồm các thành phần như:

- *Mô hình ontology CK-ONTO* mô tả tri thức của lĩnh vực, làm căn cứ để biểu diễn ngữ nghĩa cho tài liệu. Kỹ thuật xây dựng ontology cho lĩnh vực, cùng với cách thức tổ chức lưu trữ và quản lý ontology trên máy tính cũng được xem xét.

- *Các mô hình đồ thị keyphrase* biểu diễn cho nội dung của tài liệu thuộc miền và kỹ thuật xây dựng đồ thị

- *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

Nhìn chung, phương pháp mới có sự vượt trội hơn hẳn (về độ chính xác, độ bao phủ, F1 score) khi so với các phương pháp tìm kiếm theo từ khóa (VSM, BM25, LDA, MySQL FTS, RoBERTa và MPNet). Các kết quả thực nghiệm bước đầu cho thấy giải pháp đã đề xuất là khả quan và có khả năng ứng dụng tốt. Giải pháp tìm kiếm khi được triển khai thành sản phẩn ứng dụng, đã đáp ứng tốt hơn nhu cầu tìm kiếm tài liệu của người dùng. Thông qua việc thực nghiệm và đánh giá với các tiếp cận truyền thống, đề tài góp phần khẳng định được giá trị của việc khai thác thông tin ontology miền và biểu diễn văn bản bằng đồ thị vào việc giải quyết bài toán tìm kiếm. Kỹ thuật biểu diễn tài liệu giải quyết được (phần nào) hai vấn đề nhập nhằng của ngôn ngữ tự nhiên là từ đồng nghĩa và từ nhiều nghĩa. Các kỹ thuật đánh giá độ tương đồng ngữ nghĩa đạt được độ chính xác cao, dẫn đến hiệu quả tìm kiếm được cải thiện so với các

phương pháp truyền thống. Từ đó, góp thêm động lực cho hướng tiếp cận này trong các đề xuất tương lai.

Kết quả thứ hai là đề xuất một giải pháp tổng thể cho việc thiết kế và xây dựng một lớp hệ thống ứng dụng mới gọi là "Hệ quản lý cơ sở tài liệu văn bản theo ngữ nghĩa". Giải pháp tổng thể bao gồm: Đưa ra những đặc trưng cơ bản của hệ thống để phân biệt với những loại hệ thống khác; Kiến trúc hệ thống; Quy trình xây dựng; Đặt ra một số vấn đề kỹ thuật cần phải giải quyết, kinh nghiệm thực tiễn khi xây dựng các hệ thống ứng dụng.

Giải pháp đã được áp dụng và xây dựng thành công 03 hệ thống ứng dụng cụ thể, từ đó chứng minh được tính khả thi và hữu ích của những ý tưởng đã đề xuất. Giải pháp đã đáp ứng tốt hơn nhu cầu tìm kiếm tài liệu của người dùng, có thể mở rộng cho nhiều miền tri thức, nhiều ứng dụng khác nhau, và cho ngôn ngữ tiếng Việt.

Một số điểm mới nổi bật được thảo luận trong giải pháp bao gồm:

- *Mô hình cơ sở tài liệu có ngữ nghĩa (Semantic Document Base - SDB)* là một mô hình tổ chức lưu trữ và quản lý kho tài liệu trên máy tính, trong đó có hỗ trợ biểu diễn và xử lý ngữ nghĩa liên quan đến nội dung tài liệu.

- *Một phương pháp tổ chức "cơ sở về các tài liệu" theo mô hình SDB trên máy tính* (gồm các lớp mô hình lưu trữ theo hệ thống thư mục có quy chuẩn, lớp database, lớp semantic - quản lý ngữ nghĩa).

- *Các vấn đề, quy trình, kỹ thuật xử lý và tìm kiếm dựa trên các độ đo tương đồng về ngữ nghĩa.*

Kết quả thứ ba là đề xuất một phương pháp mới giải quyết bài toán Đo lường độ tương đồng ngữ nghĩa giữa hai tài liệu.

Trong đề tài này, 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 được đề xuất bao gồm:

- *Kỹ thuật rút trích keyphrase* có khai thác cơ sở tri thức DBPedia và Wikipedia.

- *Các mô hình và kỹ thuật xây dựng đồ thị keyphrase* biểu diễn cho nội dung của tài liệu thuộc tri thức tổng quát nói chung.

- *Kỹ thuật đánh giá độ tương đồng* ngữ nghĩa giữa hai tài liệu, dựa trên cách tiếp cận so khớp các đồ thị keyphrase biểu diễn chúng

Điểm mạnh của phương pháp mới là 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 rất 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.

Về mặt ứng dụng

Luận án đã nêu lên các ưu thế và lợi ích của việc nghiên cứu phát triển các mô hình cùng với các thuật giải tự động dựa trên tri thức thông qua việc thiết kế, cài đặt và xây dựng được 03 ứng dụng thử nghiệm, bao gồm:

- Đầu tiên là hệ quản lý kho tài nguyên học tập về lĩnh vực Khoa học máy tính trong phạm vi của một trường đại học, phục vụ cho các đối tượng là người học, người dạy và kể cả những người quản lý (giới hạn kho tài liệu tiếng Anh)
- Thứ hai là hệ thống hỗ trợ tìm kiếm việc làm và tuyển dụng ngành Công nghệ thông tin (giới hạn ngôn ngữ tiếng Anh)
- Cuối cùng là hệ thống hỗ trợ tìm kiếm, chọn lọc tin bài trên các báo mạng phục vụ cho nhu cầu thực tế của Phòng Báo chí và Xuất bản của Sở Thông tin và Truyền thông Bình Dương (giới hạn ở môt số lĩnh vực đặc biệt như Lao động việc làm, Đầu tư công và đầu tư nước ngoài). Đặc biệt, đây là một thử nghiệm mở rộng cho ngôn ngữ tiếng Việt.

Đóng góp khác

Các đóng góp khác của luận án bao gồm:

Với ba hệ thống ứng dụng được nghiên cứu phát triển như đã kể trên, đề tài đã xây dựng ba cơ sở tri thức tương ứng cho các lĩnh vực Khoa học máy tính, Việc làm ngành Công nghệ Thông tin, lĩnh vực Lao động – Việc làm và Đầu tư công – Đầu tư nước ngoài. Các cơ sở tri thức được lưu trữ trong cơ sở dữ liệu quan hệ và theo chuẩn OWL, được tối ưu hóa cho phép tìm kiếm. Việc quản lý cập nhật ontology được hỗ trợ bằng mô-đun quản lý với đầy đủ các tính năng cơ bản, phù hợp với yêu cầu sử dụng.

Luận án cũng bàn về các kinh nghiệm thực tiễn khi xây dựng một hệ thống tìm kiếm tài liệu theo ngữ nghĩa dựa trên SDB framework và giới thiệu chi tiết về cách tiến hành xây dựng các bộ dữ liệu thử nghiệm để đánh giá hiệu quả tìm kiếm.

HẠN CHẾ CỦA LUẬN ÁN

- Luận án không tập trung nghiên cứu vấn đề cập nhật tự động tri thức, không đi sâu vào việc suy luận giải quyết vấn đề trên tri thức, cũng như không đặt ra vấn đề phải đánh giá ontology một cách độc lập.
- Vấn đề lập chỉ mục tự động cho các tài liệu vẫn chưa được tập trung nghiên cứu chuyên sâu vì đòi hỏi phải thông qua một qui trình xử lý phức tạp trong phân tích ngữ nghĩa văn bản.
- Việc xây dựng cơ sở tri thức cho một lĩnh vực cũng gặp nhiều khó khăn vì tốn nhiều chi phí xây dựng và duy trì vốn phải có sự can thiệp của con người, đòi hỏi kiến thức của chuyên gia về lĩnh vực và phụ thuộc nhiều vào ngôn ngữ.

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".

NHỮNG KẾT QUẢ CÓ LIÊN QUAN CỦA NGHIÊN CỨU SINH

CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ

[CT1] **ThanhThuong T. Huynh**, TruongAn PhamNguyen and Nhon V. Do, "A Method for Designing Domain-Specific Document Retrieval Systems using Semantic Indexing," *International Journal of Advanced Computer Science and Applications*, ISSN 2158-107X, Vol. 10, No. 10, pp. 461-481, 2019.

[CT2] **ThanhThuong T. Huynh**, Nhon V.Do, TruongAn N.Pham, "A semantic document retrieval system with semantic search technique based on knowledge base and graph representation," in *Proceedings of The 17th International Conference on New Trends in Intelligent Software Methodologies, Tools, and Techniques*, IOS Press, 2018, pp. 870-882.

[CT3] Nhon V.Do,TruongAn PhamNguyen, Hung K. Chau, and **ThanhThuong T. Huynh**, "Improved Semantic Representation and Search Techniques in a Document Retrieval System Design," *Journal of Advances in Information Technology*, Vol. 6, No. 3, pp. 146-150, 2015.

[CT4] **ThanhThuong T. Huynh**, TruongAn PhamNguyen, and Nhon V. Do, "A Keyphrase Graph-Based Method for Document Similarity Measurement," *Engineering Letters*, Vol. 30, No. 2, pp. 692-710, 2022.

[CT5] **ThanhThuong T. Huynh**, TruongAn N.Pham, Nhon V.Do, "Keyphrase Graph in text representation for document similarity measurement," in *Proceedings of The 19th International Conference on New Trends in Intelligent Software Methodologies, Tools, and Techniques*, IOS Press, 2020, pp. 459-472.

ĐỀ TÀI NGHIÊN CỨU KHOA HỌC

[ĐT1] **Huỳnh Thị Thanh Thương**, Đỗ Văn Nhơn, Phạm Nguyễn Trường An, "Nghiên cứu phát triển một số mô hình và kỹ thuật trong việc thiết kế, xây dựng hệ quản lý kho tài liệu văn bản theo ngữ nghĩa", Đề tài cấp ĐHQG-HCM loại C, thời gian thực hiện từ tháng 04/2018 đến tháng 11/2019 (Chủ nhiệm đề tài: Huỳnh Thị Thanh Thương).

[ĐT2] **Huỳnh Thị Thanh Thương**, Đỗ Văn Nhơn, Phạm Nguyễn Trường An, Châu Kim Hùng, "Nghiên cứu phát triển giải pháp quản trị kho tài nguyên học tập theo ngữ nghĩa", Đề tài cấp cơ sở, thời gian thực hiện từ tháng 07/2014 đến tháng 12/2015 (Chủ nhiệm đề tài: Huỳnh Thị Thanh Thương).