

Một phương pháp tìm kiếm dựa trên Ontology

phục vụ quản lý thông tin khoa học công nghệ

Trần Đình Khang, Vũ Tuyết Trinh, Đỗ Đức Thành, Đỗ Thị Ngọc Quỳnh

Bộ môn Hệ thống Thông tin

Đại học Bách Khoa Hà Nội

{khangtd-fit, trinhvt-fit, thanhdd-fit, quynhdtm-fit}@mail.hut.edu.vn

Tóm tắt

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. Trong lĩnh vực khoa học công nghệ nói riêng và các lĩnh vực khác nói chung, khi khối lượng thông tin, tài liệu văn bản ngày càng lớn, vấn đề tìm kiếm thông tin dẫn xuất, tìm kiếm theo ngữ nghĩa là rất cần thiết trong việc phát hiện những tri thức bổ sung. Trong bài báo này, chúng tôi trình bày một phương pháp tìm kiếm tài liệu, dữ liệu dựa trên ontology, phục vụ cho việc quản lý tài liệu và thông tin trong lĩnh vực khoa học công nghệ. Phương pháp tìm kiếm được ứng dụng trong một hệ thống quản lý tài liệu điện tử và thông tin khoa học công nghệ. Các thử nghiệm cho thấy phương pháp tìm kiếm dữ liệu dựa trên ontology có khả năng phát hiện các tri thức bổ sung tốt hơn so với các phương pháp tìm kiếm thông thường.

1. Đặt vấn đề

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. Trong lĩnh vực khoa học công nghệ nói riêng và các lĩnh vực khác nói chung, khi khối lượng thông tin, tài liệu văn bản ngày càng lớn, vấn đề tìm kiếm thông tin dẫn xuất, tìm kiếm theo ngữ nghĩa là rất cần thiết trong việc phát hiện những tri thức bổ sung.

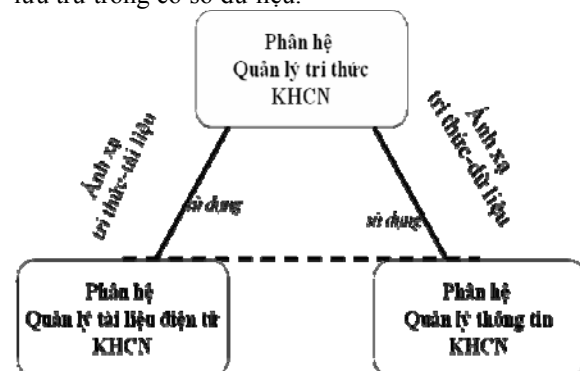
Mục đích của chúng tôi là xây dựng một hệ thống quản lý và lưu trữ thông tin khoa học công nghệ với khả năng tìm kiếm dựa trên ontology. Hệ thống này không chỉ hỗ trợ tìm kiếm dựa trên từ khóa và tìm kiếm trên cấu trúc dữ liệu lưu trữ mà còn hỗ trợ tìm kiếm dựa trên tri thức của lĩnh vực. Với mục đích đó, chúng tôi đề xuất một ontology cho lĩnh vực khoa học

công nghệ và khai thác các suy diễn ngữ nghĩa trên ontology này để phục vụ các tìm kiếm xấp xỉ, tìm kiếm dựa trên ngữ nghĩa trong hệ thống quản lý tài liệu điện tử và thông tin khoa học công nghệ.

Phần tiếp theo của báo cáo được tổ chức như sau: Mục 2 trình bày tổng quan về cách tiếp cận của chúng tôi trong việc xây dựng hệ quản trị tài liệu với khả năng tìm kiếm dựa trên ngữ nghĩa. Mục 3 giới thiệu quy trình xây dựng ontology khoa học. Mục 4 tập trung giới thiệu các kỹ thuật cơ bản hỗ trợ tìm kiếm ngữ nghĩa cho hệ thống đề cập trong báo cáo này. Mục 5 đề cập đến một vài vấn đề cơ bản trong triển khai và thử nghiệm hệ thống. Cuối cùng, một số ý kiến trao đổi, đánh giá và so sánh với các nghiên cứu có liên quan sẽ được trình bày trong mục 6.

2. Các tiếp cận xây dựng mô tơ tìm kiếm

Phạm vi thông tin tìm kiếm trong hệ thống không chỉ bao gồm các tài liệu điện tử về lĩnh vực khoa học và công nghệ mà còn cả các thông tin có cấu trúc được lưu trữ trong cơ sở dữ liệu.



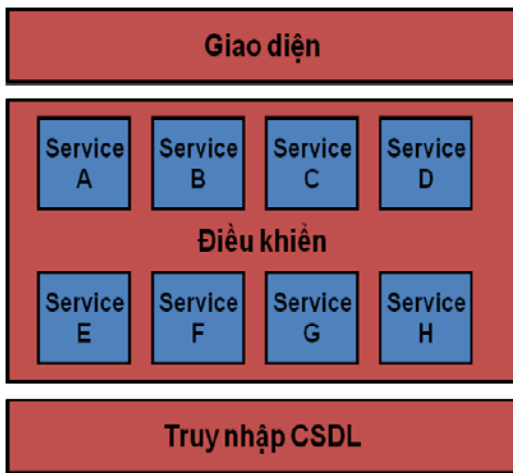
Hình 1 Cách tiếp cận

Hệ thống được xây dựng với ba phân hệ chính như chỉ ra trong hình 1.

- (i) Phân hệ quản lý tài liệu điện tử có các chức năng cho phép lưu trữ, quản lý quy trình nghiệp vụ xử lý và thao tác với tài liệu.
- (ii) Phân hệ quản lý thông tin khoa học công nghệ cho phép lưu trữ và quản lý các dữ liệu về đề tài, sản phẩm, các chuyên gia và đơn vị trong lĩnh vực khoa học công nghệ.
- (iii) phân hệ quản lý tri thức khoa học công nghệ tạo và quản lý tri thức trong lĩnh vực khoa học công nghệ. Một chức năng quan trọng của phân hệ này là suy diễn. Dựa trên kết quả suy diễn này, hệ thống hỗ trợ khả năng tìm kiếm mở rộng dựa trên ngữ nghĩa và tri thức

Việc phân chia hệ thống thành 3 phân hệ cho phép phát triển và triển khai hệ thống một cách dễ dàng và thuận tiện hơn. Tùy theo yêu cầu cụ thể của đơn vị ứng dụng, một hay nhiều phân hệ có thể được cài đặt với cấu hình phù hợp.

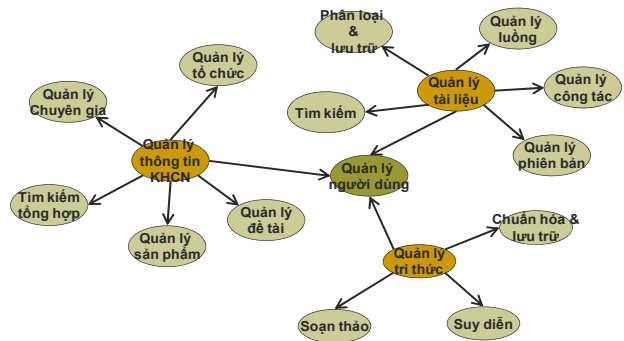
Dựa trên cách tiếp cận này, chúng tôi lựa chọn kiến trúc hướng dịch vụ để xây dựng hệ thống BKDoST như chỉ ra trong hình 2.



Hình 2 Mô hình hệ thống

Theo mô hình này, các chức năng hệ thống được xây dựng dưới dạng các dịch vụ (*services*) tương tác với nhau thông qua giao diện của dịch vụ để đảm bảo các chức năng của hệ thống. Với cách tiếp cận này, chúng tôi đã lựa chọn JSF và Spring Framework như nền tảng cho việc xây dựng hệ thống.

Hình 3 giới thiệu tổng quan về hệ thống và làm rõ mối liên kết giữa các phân hệ trong hệ thống.



Hình 3 Các chức năng của hệ thống

Qua hình 3 chúng ta có thể nhận thấy sự độc lập tương đối giữa ba phân hệ của hệ thống.

- Các chức năng quản lý đề tài, quản lý chuyên gia, quản lý sản phẩm và quản lý đơn vị khoa học công nghệ cung cấp các tiện ích thực hiện thêm, sửa, xóa dữ liệu tương ứng trong cơ sở dữ liệu cài đặt bởi MySQL.
- Các chức năng phân loại và lưu trữ tài liệu, quản lý luồng công việc và phiên bản tài liệu, quản lý công tác hỗ trợ quá trình thao tác đồng thời với tài liệu của nhiều người dùng. Việc quản lý lưu trữ và thao tác với tài liệu được xây dựng trên nền của phần mềm mã nguồn mở Alfresco.
- Soạn thảo, chuẩn hóa và suy diễn chịu trách nhiệm định nghĩa, chia sẻ và khai thác các tri thức trong lĩnh vực giữa nhiều người dùng khác nhau.
- Việc tích hợp hệ thống được thực hiện thông qua (i) tích hợp mô-đun quản lý người dùng. Người dùng hệ thống đăng nhập một lần và có thể khai thác tất cả các chức năng của các phân hệ khác nhau tùy theo phân quyền của họ; (ii) giao diện thống nhất giữa các chức năng tìm kiếm dựa trên dữ liệu, tìm kiếm tài liệu với khả năng suy diễn dựa trên tri thức.

3. Xây dựng ontology khoa học công nghệ

Để có thể tìm kiếm được các kết quả dẫn xuất, thông tin phải được biểu diễn theo khuôn dạng giúp cho máy tính hiểu và thông dịch được ngữ nghĩa của thông tin đó. Trong lĩnh vực khoa học công nghệ, hàng năm có rất nhiều các đề tài, công trình nghiên cứu hay sản phẩm thuộc các cấp khác nhau. Tuy nhiên việc tìm kiếm sử dụng các kết quả có sẵn không phải là đơn giản do không có hệ thống quản lý chung hoặc có nhưng các hệ thống này không tương thích. Xuất phát từ nhu cầu đó, việc xây dựng một Ontology về khoa học công nghệ là hết sức cần thiết. Ontology về khoa học công nghệ sẽ cung cấp cho người dùng biết được

nhiều thông tin bổ ích, như: tác giả, cơ quan chủ quản, tóm tắt, toàn văn của đề tài, công trình nghiên cứu hay các thông tin về sản phẩm, người liên hệ và nhiều thông tin khác.

Một cách hình thức, có thể hiểu khái niệm ontology là một tập định nghĩa của các khái niệm cơ bản mà máy tính có thể hiểu được trong một vài lĩnh vực nào đó và các mối liên hệ giữa các khái niệm từ đó có thể trích rút tri thức. Ontology không chỉ là một bảng từ vựng phù hợp: Ontology cung cấp nền tảng vững chắc cho việc xây dựng hệ quản lý tri thức ở mức độ cao, các thuật ngữ trong ontology được lựa chọn để đảm bảo rằng hầu hết các khái niệm cơ bản và sự khác biệt được định nghĩa và chỉ rõ. Ontology không chỉ là sự phân cấp các thuật ngữ: Mặc dù sự phân cấp các thuật ngữ đóng góp ngữ nghĩa cho các thuật ngữ trong bảng từ vựng, ontology bao gồm nhiều mối quan hệ giữa các thuật ngữ, những mối quan hệ này cho phép biểu diễn tri thức miền mà không cần sử dụng các thuật ngữ biểu diễn tri thức miền.

Quá trình xây dựng ontology cho một lĩnh vực thông thường tuân theo các bước sau: (1) xác định miền và phạm vi của ontology; (2) định nghĩa lĩnh vực và phạm vi của ontology; (3) định nghĩa các khái niệm – xây dựng Tbox; (4) tạo các cá thể - xây dựng bộ Abox. Phần tiếp theo chúng tôi quá trình xây dựng Ontology trong lĩnh vực khoa học công nghệ tuân theo bốn bước trên.

3.1. Xác định miền và phạm vi của ontology

Miền mà Ontology KHCN sẽ bao trùm là khái niệm, thông tin và đánh giá của các đề tài, tài liệu, sản phẩm, văn bản và các công trình khoa học. Chúng ta sẽ dùng Ontology KHCN để tra cứu các đề tài, tra cứu các sản phẩm công nghệ, tìm kiếm chuyên gia, tìm kiếm tài liệu, giải pháp, công nghệ... Người bảo trì Ontology KHCN có thể chính là tác giả, cùng với toàn bộ người dùng quan tâm đến KHCN và có những hiểu biết nhất định về Ontology sẽ nâng cấp thông tin khi có thay đổi. Ontology KHCN có thể trả lời được các câu hỏi tiềm tàng có dạng như: Có những đề tài nào thuộc lĩnh vực mà người dùng quan tâm? Đề tài nào dành được sự quan tâm nhiều nhất cũng như nhận định về giá trị, khả năng ứng dụng vào thực tiễn? Tài liệu đang được xem xét có những phiên bản nào, sự đánh giá của các độc giả đối với các phiên bản của tài liệu này như thế nào? Tìm những chuyên gia đa lĩnh vực: ví dụ chuyên gia vừa trong lĩnh vực CNTT vừa trong lĩnh vực Hoá sinh?

3.2. Định nghĩa lĩnh vực và phạm vi của OntologyKHCN

Lĩnh vực mà chúng ta cần xây dựng ontology là thông tin liên quan đến khoa học công nghệ, mà cụ thể ta xem xét các thông tin liên quan đến thông tin đề tài, sản phẩm, quy trình công nghệ, tài liệu khoa học, văn bản, tin tức. Dựa trên quá trình khảo sát nhu cầu quản lý thông tin tại phòng KHCN thuộc Đại học Bách Khoa Hà Nội, tại phòng KHCN thuộc sở Khoa học Công nghệ Thành Phố Hà Nội, tại sở Bưu chính Viễn thông, chúng tôi đã xây dựng một số khái niệm liên quan đến khoa học công nghệ, được trình bày trong các phần tiếp sau.

3.3. Định nghĩa các khái niệm – Xây dựng Tbox

Để biểu diễn tri thức về khoa học công nghệ công việc trước tiên ta phải làm đó là xây dựng các khái niệm khoa học công nghệ từ các khái niệm nguyên thủy, các quan hệ nguyên thủy và các khái niệm mở rộng. Hệ thống khái niệm mà ta có được gọi là bộ thuật ngữ (TBox). Đây là một trong hai thành phần chính của hệ cơ sở tri thức dựa vào logic mô tả.

Đầu tiên, chúng tôi định nghĩa các khái niệm nguyên thủy, bao gồm các khái niệm về: chuyên gia, đề tài, sản phẩm, tiêu chí đánh giá, các đơn vị, lĩnh vực...

Tiếp đó, dựa trên các khái niệm nguyên thủy, chúng tôi định nghĩa các quan hệ nguyên thủy giữa chúng. Một số quan hệ nguyên thủy cơ bản bao gồm: là chủ nhiệm đề tài, là chuyên gia thuộc lĩnh vực, có tham gia đề tài thuộc lĩnh vực, có sản phẩm thuộc lĩnh vực... Ví dụ: trung tâm an ninh mạng BKIS có sản phẩm là phần mềm diệt virus BKAV.

Sau khi định nghĩa các khái niệm nguyên thủy và các quan hệ nguyên thủy, chúng tôi tiến hành định nghĩa các khái niệm mở rộng. Ví dụ, một chuyên gia trong lĩnh vực công nghệ thông tin là chuyên gia thuộc lĩnh vực công nghệ thông tin hoặc là chủ nhiệm ít nhất một đề tài thuộc lĩnh vực công nghệ thông tin, hoặc tham gia ít nhất ba đề tài thuộc lĩnh vực công nghệ thông tin. Hoặc một chuyên gia Hóa sinh là người vừa là chuyên gia trong lĩnh vực hóa học, lại vừa là chuyên gia trong lĩnh vực sinh học.

3.4. Tạo các cá thể - Xây dựng bộ Abox

Ngoài bộ thuật ngữ TBox vừa trình bày, thành phần thứ hai của cơ sở tri thức là bộ khẳng định ABox. Bằng bộ khẳng định người ta biểu diễn các cá thể. Ta ký hiệu các cá thể là những ký tự a, b, c. Dùng các

chứa thông tin về các tài liệu văn bản khoa học công nghệ như văn bản pháp quy, thuyết minh đề tài...

4.3. Module sản sinh cá thể

Module này xử lý trên toàn bộ dữ liệu để tìm ra các cá thể phù hợp với các danh từ và động từ mà chúng ta đã xây dựng. Tập cá thể này sẽ được lưu lại để sử dụng trong quá trình tìm kiếm. Thông tin lưu lại đủ để lần lại cá thể thực sự trong quá trình tìm kiếm. Mỗi cá thể tìm được chính là một thực thể thuộc về một khái niệm nào đó trong ontology.

Chúng tôi không lưu trữ thông tin về các cá thể trong cùng một tệp ontology template bởi lý do, nếu số lượng cá thể thoải mãn lớn, việc lưu trữ tập trung các định nghĩa và các cá thể trong cùng một tệp sẽ làm cho việc suy diễn trở nên khó khăn và không hiệu quả. Thay vào đó, các cá thể được lưu trữ trong một kho lưu trữ cá thể riêng biệt. Giải pháp Instance Store được sử dụng để lưu trữ các cá thể này. InstanceStore đã có sẵn cơ chế để lưu các cá thể dựa theo định nghĩa ontology ban đầu cho nên nhiệm vụ của module trích rút lúc này chỉ còn là tìm kiếm các cá thể sau đó sử dụng instance store để lưu trữ các cá thể này. Thông thường, mỗi cá thể kèm theo Id của chúng ở trong hệ cơ sở dữ liệu đã có để sau này module tìm kiếm có thể tìm lại được.

Để giảm bớt số lượng các suy diễn phải thực hiện tại thời điểm có truy vấn của người dùng, Instance store cố gắng lưu trữ càng nhiều càng tốt các mô tả (hay thông tin) về cá thể vào trong cơ sở dữ liệu. Các mô tả này được rút ra từ các khẳng định do người dùng nhập vào hoặc được rút ra ngay trong quá trình suy diễn để trả lời các truy vấn trước đó của người dùng.

4.4. Module suy diễn

Module suy diễn có nhiệm vụ nhận yêu cầu tìm kiếm của người dùng, tiến hành suy diễn trên cơ sở tri thức của hệ thống, sau đó hiển thị kết quả trả về. Thuật toán suy diễn Tableau được sử dụng để cài đặt module này. Chúng tôi sử dụng các API của Pellet [] để xây dựng module này. Pellet, một trong các chương trình suy diễn được sử dụng phổ biến nhất hiện nay, được xây dựng theo một kiến trúc cho phép người dùng có thể triển khai một thuật toán tableau cho một họ ngôn ngữ logic mô tả mới sau đó ghép vào Pellet.

Module này thực hiện các công việc liên quan đến suy diễn sau:

- Kiểm tra tính nhất quán: Đảm bảo rằng một ontology không chứa các mâu thuẫn. Trong logic mô tả đây là vấn đề kiểm tra xem Abox có nhất quán không dựa theo một Tbox.

- Kiểm tra tính thỏa của khái niệm: Kiểm tra xem liệu một khái niệm có thể có các thể hiện được hay không. Nếu một khái niệm mà không có tính thỏa thì khi ta thêm một thể hiện của khái niệm đó, toàn bộ ontology của ta sẽ không còn tính nhất quán.
- Phân loại: Xác định mối quan hệ giữa tất cả các lớp (khái niệm) trong ontology qua đó xây dựng nên cây phân cấp lớp của ontology. Cây phân cấp này sau đó có thể được dùng để trả lời các câu truy vấn như tìm kiếm tất cả các con trực tiếp của một lớp.
- Xác định thể hiện: Xác định lớp thấp nhất (trong cây phân cấp, tức là lớp ít trừu tượng nhất) mà một thể hiện thuộc vào. Nói một cách khác, khả năng này cho phép ta xác định khái niệm cho tất cả các thể hiện ở trong Abox.

5. Thử nghiệm và đánh giá

5.1. Môi trường thử nghiệm

Hệ thống tìm kiếm dựa trên ontology được thử nghiệm trên một máy tính Pentium IV 3.0 GHz, 480MB RAM. Cơ sở dữ liệu thông tin khoa học công nghệ được lưu trữ bằng hệ quản trị CSDL MySQL 5.0.45. Cơ sở dữ liệu này chứa dữ liệu về khoảng 3000 chuyên gia, 1500 đề tài cùng với hơn 150 lĩnh vực KHCN.

Ontology Khoa học Công nghệ được xây dựng bởi phần mềm soạn thảo cơ sở tri thức. Phần mềm này được viết dựa trên các API của Protégé. Ontology KHCN định nghĩa các khái niệm và thuộc tính liên quan đến đề tài, sản phẩm, chuyên gia khoa học công nghệ. Ví dụ, khái niệm một chuyên gia trong lĩnh vực công nghệ thông tin được mô tả dưới dạng OWL DL như sau:

```
intersectionOf(
restriction(laChuNhiemDeTai Chuyen_gia
someValuesFrom
De_tai)
restriction(coTuKhoa someValuesFrom Tin_hoc) )
```

Với mô tả ở trên, Chuyên gia công nghệ thông tin là những chuyên gia:

- Có lĩnh vực chuyên môn là công nghệ thông tin
- Hoặc là chủ nhiệm đề tài thuộc lĩnh vực công nghệ thông tin
- Hoặc có tham gia nhiều hơn 3 đề tài công nghệ thông tin

Trong CSDL chuyên gia KHCN, lĩnh vực chuyên môn của chuyên gia được phản ánh trong trường lĩnh vực. Chúng tôi định nghĩa những khái niệm mở rộng

không được phản ánh trong các trường này. Ví dụ, khái niệm một chuyên gia hoá sinh (vừa là chuyên gia trong lĩnh vực hoá học, vừa là chuyên gia trong lĩnh vực sinh học) được định nghĩa dưới dạng OWL-DL như sau:

```

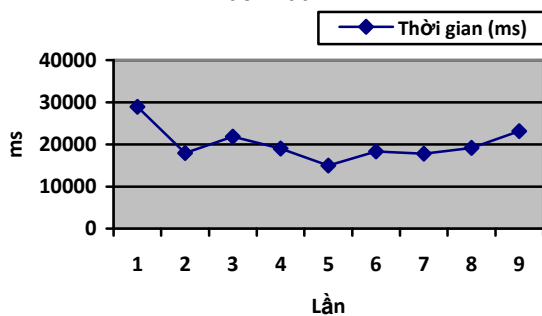
OWL DL : intersectionOf(
  Chuyen_gia
  restriction(laChuNhiemDeTai someValuesFrom
De_tai)
  restriction(coTuKhoa
someValuesFrom Hoa_hoc)
  restriction( coTuKhoa someValuesFrom Sinh
hoc))

```

Tương tự như trên, chúng tôi định nghĩa một loạt những khái niệm mở rộng khác như: chuyên gia hóa lý, chuyên gia tin sinh học... Sự xuất hiện của các khái niệm mở rộng trong ontology sẽ giúp tìm kiếm ra những khái niệm dẫn xuất mà khi không thể tìm thấy khi thực hiện với phương pháp truy vấn trên CSDL quan hệ thông thường.

Sau khi đã định nghĩa các khái niệm cơ sở cũng như các khái niệm mở rộng, chúng tôi tiến hành sinh các thể hiện của CSTT. Mô đun sản sinh thể hiện đọc các định nghĩa trong ontology template, quét duyệt toàn bộ CSDL quan hệ, tìm những thể hiện thỏa mãn định nghĩa. Chúng tôi đã tiến hành đo thời gian trung bình của quá trình sinh các thể hiện này trên hệ thống máy tính thử nghiệm. Kết quả đo được minh họa trong hình 8. Kết quả thử nghiệm cho thấy, với cấu hình máy tính hiện tại, hệ thống mất khoảng 2s để sản sinh khoảng 3000 thể hiện. Đây là một khoảng thời gian chấp nhận được.

Thời gian sinh thể nghiệm trong các lần khác nhau



Hình 6 Thời gian sinh các thể hiện

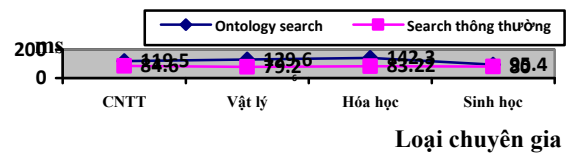
5.2. Độ đo sử dụng

Dựa trên CSTT đã xây dựng, chúng tôi so sánh phương pháp tìm kiếm dựa trên ontology với phương pháp tìm kiếm thông thường, dựa trên các truy vấn SQL đến CSDL quan hệ. Để tiến hành so sánh, chúng tôi sử dụng 2 độ đo.

Độ đo thứ nhất là thời gian tìm kiếm, ký hiệu T, được tính từ thời điểm gửi đi câu truy vấn đến thời điểm trả về kết quả. Độ đo thứ hai là số lượng kết quả đúng trả về, ký hiệu TP (True Positive). Chúng ta kỳ vọng TP của phương pháp tìm kiếm dựa trên ontology sẽ lớn hơn hoặc bằng TP của phương pháp tìm kiếm thông thường, đồng thời thời gian tìm kiếm của phương pháp dựa trên ontology là chấp nhận được.

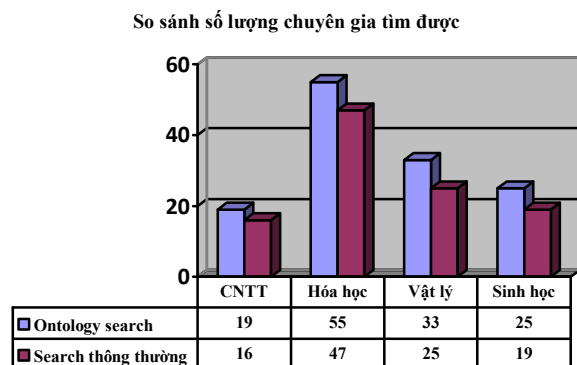
5.3. Case study 1: tìm kiếm các khái niệm cơ bản

Chúng tôi sử dụng hệ thống tìm kiếm dựa trên ontology để tìm kiếm tất cả chuyên gia trong một lĩnh vực nào đó, so sánh kết quả tìm được với tìm kiếm đơn giản. Đơn giản ở đây tức là liệt kê ra tất cả các chuyên gia thuộc lĩnh vực đó dựa vào trường lĩnh vực của chuyên gia trong cơ sở dữ liệu.



So sánh tốc độ giữa 2 kiểu tìm kiếm

Hình 7 So sánh thời gian tìm kiếm của các phương pháp

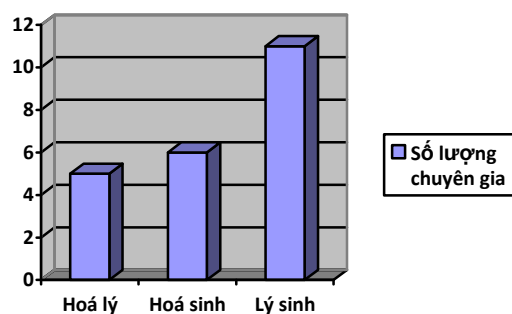


Hình 8 So sánh kết quả tìm kiếm đúng của các phương pháp

Chúng tôi tiến hành tìm kiếm chuyên gia trong các lĩnh vực CNTT, vật lý, hoá học và sinh học. Kết quả thử nghiệm được minh họa trong hình 7 và 8. Trong bốn lần thử nghiệm, phương pháp tìm kiếm ontology có T lớn hơn so với tìm kiếm truy vấn trên CSDL. Tuy nhiên, TP của nó lại lớn hơn. Ví dụ, để tìm kiếm các chuyên gia CNTT, tìm kiếm trên ontology phải mất gần 120ms và trả về 19 bản ghi đúng. Trong khi đó tìm kiếm bằng truy vấn CSDL thông thường chỉ mất 85ms nhưng chỉ trả về được 16 bản ghi đúng. Điều này là do khái niệm chuyên gia CNTT trong CSTT được định nghĩa không chỉ là các chuyên gia có trường lĩnh vực trong CSDL là CNTT, mà còn là những người chủ trì ít nhất 1 đề tài thuộc lĩnh vực CNTT hoặc tham gia ít nhất ba đề tài CNTT. Rõ ràng là tìm kiếm dựa trên ontology sẽ giúp tìm ra số lượng kết quả lớn hơn.

5.4. Case study 2: Tìm kiếm các khái niệm dẫn xuất

Trong thử nghiệm thứ hai, chúng tôi tiến hành tìm kiếm với một số định nghĩa đặc biệt trong ontology, cụ thể là lần lượt tìm kiếm các chuyên gia hoá lý, hoá sinh, và lý sinh. Ví dụ, chuyên gia hóa lý là người vừa là chuyên gia hóa học vừa là chuyên gia vật lý. Tìm kiếm bằng truy vấn SQL đến CSDL quan hệ không thể tìm ra các khái niệm này. Với phương pháp tìm kiếm dựa trên ontology, hệ thống đã tìm ra được 5 chuyên gia hoá sinh, 6 chuyên gia hoá lý, và 11 chuyên gia hoá sinh. Kết quả thử nghiệm được minh họa trong hình 9.



Hình 9 Kết quả tìm kiếm các khái niệm dẫn xuất

6. Thảo luận

6.1. Quản lý tài liệu

Để quản lý vòng đời của các tài liệu, ta có thể sử dụng một hệ thống quản trị tài liệu. Một hệ thống quản trị tài liệu được thiết kế tốt phải là một hệ thống cho phép dễ dàng tìm kiếm và chia sẻ thông tin. Trong phạm vi nghiên cứu của mình chúng tôi sử dụng một hệ thống quản trị tài liệu mở. Dựa trên một số tính năng sẵn có chúng tôi cải tiến, thay đổi mô hình hệ thống để phù hợp với nghiên cứu của mình. Hiện nay có rất nhiều hệ thống quản trị tài liệu mở như DotNetNuke[1], Joomla[2], Drupal[3], Plone[4] hay Alfresco [8].

Alfresco có thể một số lượng lớn các tài liệu điện tử trong các “không gian” mềm dẻo, thông minh. Các tài liệu có thể được truy nhập thông qua một giao diện web, các mạng thư mục chia sẻ tài liệu, FTP, WebDav, và các phương pháp khác. Người dùng có thể sử dụng Alfresco để xử lý tài liệu theo các quy luật và các workflow. Hệ thống cũng cho phép áp dụng các chức năng quản lý phiên bản cho tài liệu một cách tự động. Không những thế, Alfresco có khả năng triển khai theo nhiều mô hình khác nhau từ mô hình triển khai trên một máy chủ cho đến nhiều máy chủ. Alfresco được biết đến như là một trong số những hệ thống quản trị tài liệu mã nguồn mở mạnh nhất hiện nay. Một số nghiên cứu chỉ ra rằng Alfresco cung cấp nhiều chức năng hơn các hệ thống mã nguồn mở khác [5]. Trong khi DotNetNuke sử dụng VB.Net, ASP.NET, SQL Server; Joomla, Drupal, sử dụng PHP, Alfresco dựa trên JSR-170, Spring và JSF frameworks, tương tác với các portal JSR-168, và có thể mở rộng dễ dàng. Với tất cả những ưu điểm trên chúng tôi sử dụng Alfresco cho nghiên cứu của mình

6.2. Tìm kiếm dựa trên ontology

Trong bài báo này, chúng tôi giới thiệu một phương pháp tìm kiếm dựa trên ontology cho hệ thống quản trị tài liệu khoa học công nghệ. Có rất nhiều nghiên cứu về việc sử dụng ontology trong tìm kiếm đã được tiến hành. Trong số đó có các nghiên cứu về việc xây dựng các search engine tìm kiếm các tài liệu web. Các search engine hiện nay hầu hết đều tìm kiếm dựa trên các từ khóa và các phép toán boolean. Nguyên tắc tìm kiếm chỉ đơn giản là nếu như tài liệu có các từ khóa thỏa mãn các phép toán boolean thỏa mãn câu truy vấn thì tài liệu đó là phù hợp. Việc nhúng một miền tri thức xác định vào search engine còn chưa được thực hiện rộng rãi. Hướng tiếp cận ontology sẽ cung cấp khả năng tìm kiếm ngữ nghĩa cho search engine. He Hu và Xiaoyong Du [6] cung cấp một framework tìm kiếm các tài liệu web sử dụng ontology và các luật để xử lý các vấn đề đồng nghĩa và đa nghĩa. Trong khi hướng tiếp cận của họ là xây dựng một framework dựa trên ontology cho tìm kiếm web, chúng tôi xây dựng một search engine dựa trên ontology để tìm kiếm thông tin trong một lĩnh vực xác định là khoa học công nghệ. Ngoài ra còn có rất nhiều nghiên cứu chẳng hạn như Knarig Arabshian và Henning Schulzrinne [7], họ đưa ra giải pháp kết hợp giữa các truy vấn ontology và các tìm kiếm dựa trên từ khóa. Các câu truy vấn dựa trên từ khóa sẽ được thay đổi dựa vào ontology. Hệ thống của chúng tôi đi theo một hướng tiếp cận khác. Chúng tôi dựa trên ontology để đưa ra những câu trả lời mà những tìm kiếm thông thường không thể có trong khi không thay đổi câu truy vấn của người dùng

7. Lời cảm ơn

Các kết quả thực nghiệm trong bài báo được thực hiện tại phòng thí nghiệm chuyên đề thuộc Bộ môn Hệ thống Thông tin, khoa CNTT, Đại học Bách Khoa Hà Nội. Kết quả nghiên cứu được hỗ trợ kinh phí một phần bởi dự án CNTT/03-2007-2.

8. Tài liệu tham khảo

- [1] www.dotnetnuke.com/
- [2] www.joomla.org/
- [3] <http://drupal.org/>
- [4] <http://plone.org/>
- [5] http://www.infoworld.com/article/07/10/08/41TC-open-source-cms_1.html
- [6] Adopting Ontologies and Rules in Web Searching Services. He Hu and Xiaoyong Du. Information School, Renmin University of China, Beijing China 100872.
- [7] XSEarch: A Semantic Search Engine for XML. Sara Cohen, Jonathan Mamou, Yaron Kanza, Yehoshua Sagiv. Proceedings