HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN VĂN NHÂN

MÔ HÌNH HOÁ TÀI NGUYÊN THÔNG TIN TRƯỜNG ĐẠI HỌC VÀ HỖ TRỢ TRUY XUẤT THÔNG TIN THEO TIẾP CẬN LINKED DATA

Chuyên ngành : **HỆ THỐNG THÔNG TIN**

Mã số : **8.48.01.04**

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2020

Luận văn được hoàn thành tại:
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS. HOÀNG HỮU HẠNH

Phản biện 1: PGS.TS. Nguyễn Hà Nam

Phản biện 2: TS. Nguyễn Vĩnh An

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 10 giờ 30 ngày 20 tháng 06 năm 2020

Có thể tìm hiểu luận văn tại:
- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

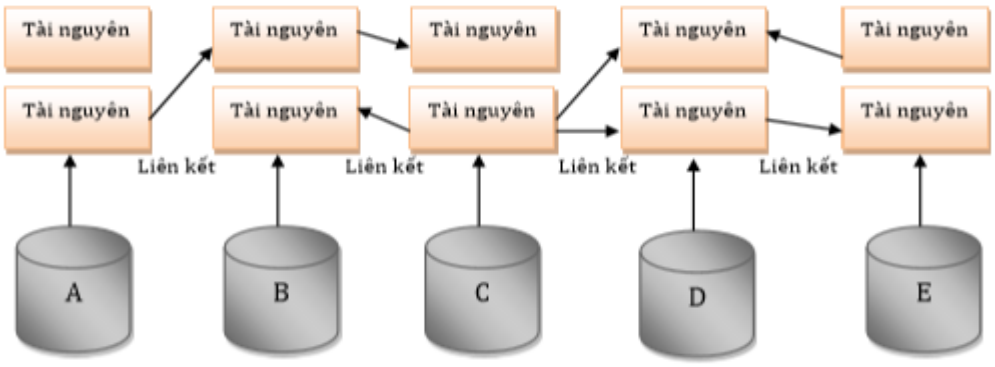
MỞ ĐẦU

1. Lý do chọn đề tài

Thuật ngữ "Linked Data" được Tim Berner-Lee đưa ra trong các ghi chép về kiến trúc "Linked Data Web" của mình. Thuật ngữ này chỉ cách thức để xuất bản và liên kết các dữ liệu có cấu trúc trên Web. Giả thuyết cơ bản của Linked Data là lợi ích và giá trị của dữ liệu tăng lên khi nó được liên kết với các dữ liệu khác. Nói cách khác, Linked Data đơn giản là sử dụng Web để tạo ra các liên kết định kiểu (typed link) giữa các dữ liệu từ nhiều nguồn tài nguyên khác nhau. Điều này giống như hai CSDL của hai tổ chức ở các vùng địa lý khác nhau hay đơn giản là một hệ thống hỗn tạp của cùng một tổ chức không dễ dàng để trao đổi, liên thông ở mức dữ liệu. Do vậy, Linked Data nhằm vào dữ liệu được đưa lên Web theo cách máy tính có thể đọc được, có ngữ nghĩa rõ ràng và nó được liên kết đến tập dữ liệu bên ngoài và ngược lại các dữ liệu đó cũng liên kết đến nó.

Linked Data sử dụng công nghệ Web hiện tại để kết nối các tài nguyên (đối tượng dữ liệu) liên quan đến nhau mà không cần liên kết trước, nghĩa là bỏ đi rào chắn liên kết dữ liệu mà hiện tại đang liên kết bởi nhiều phương thức khác nhau. Linked Data là kết nối dữ liệu phân tán trên Web.

Linked Data là thuật ngữ dùng để mô tả cách thức tốt nhất được đề nghị để duyệt, chia sẻ và kết nối các dữ liệu, thông tin, tri thức của Web ngữ nghĩa bằng cách sử dụng URIs và RDF.



Hình 1. Mô hình liên kết dữ liệu trong Web ngữ nghĩa

2. Tổng quan về vấn đề nghiên cứu

Đề tài là sự ứng dụng những đặc điểm và tính chất của mô hình và kỹ thuật Linked Data để triển khai và hoàn thiện một ứng dụng có tính áp dụng thực tiễn cao khi giải quyết được vấn đề còn tồn tại và hoa hụt do các phương thức truyền thống hay các ứng dụng hiện tại chưa đáp ứng được. Đề tài theo hướng ứng dụng và áp dụng những kỹ thuật và các công cu hỗ trợ để đề tài hoàn thiện tốt hơn.

3. Mục đích nghiên cứu

Tiếp cận và tìm hiểu về Web ngữ ngĩa (Semantic Web), các công nghệ Web ngữ nghĩa (RDF/RDFS, ontology, OWL, SPARQL), Linked Data trong các ứng dụng thông minh trong hiện nay.

Xây dựng Ứng dụng Quản lý tài nguyên thông minh theo Linked Data và hỗ trợ giảng viên và sinh viên trong tìm kiếm các tài nguyên, và dữ liệu liên quan; tiến hành phát triển demo được sản phẩm và định hướng tính ứng dụng của sản phẩm vào thực tiễn.

4. Đối tượng và phạm vi nghiên cứu

1. Đối tượng nghiên cứu:

    + Web ngữ nghĩa, RDF/RDFS, Ngôn ngữ truy vấn SPARQL

    + Mô hình dữ liệu Linked Data.

    + OWL - Web Ontology Language

2. Phạm vi nghiên cứu:

    + Đề tài thực hiện là một đề tài hướng ứng dụng và được thử nghiệm sử dụng trong môi trường nghiên cứu hoặc các cơ sở giáo dục.

5. Phương pháp nghiên cứu:

3. Phương pháp phân loại và hệ thống hoá lý thuyết
4. Phương pháp phân tích và tổng hợp lý thuyết
5. Phương pháp chuyên gia
6. Phương pháp thực nghiệm xây dượng ứng dụng theo quy trình.

Chương 1 - TỔNG QUAN VỀ WEB NGỮ NGHIÃ VÀ LINKED DATA

1 Tổng quan về Web ngữ nghĩa

1.1 Web của ngày hôm nay

World Wide Web đã thay đổi xã hội con người vô cung to lớn. Như đã thay đổi cách thức con người trao đổi với nhau cũng như cách điều hành công việc và kinh doanh và sự thay đổi là ở trung tâm của một cuộc cách mạng: chuyển đổi thế giới phát triển sang một nền kinh tế tri thức, hay nói rộng hơn là một xã hội tri thức. Sự phát triển này cũng thay đổi cách chúng ta nghĩ về các chiếc máy tính. Bây giờ máy tính không những được dùng để thực hiện các phép tính toán số học mà chúng hầu như được sử dụng cho việc xử lý thông tin, các ứng dụng đặc thù là các CSDL, xử lý văn bản, bảng tính và trò chơi điện tử.

Những thông tin hiện nay trên World Wide Web chủ yếu được biểu diễn ở dạng HTML, một ngôn ngữ phổ dụng để trình diễn thông tin. XML ra đời và trở thành một công cụ trao đổi dữ liệu không có cấu trúc, bán cấu trúc và có cấu trúc giữa các hệ thống, nâng cao sự tích hợp của các ứng dụng. Tuy nhiên, các giải pháp dựa trên XML cho quá trình tích hợp của các ứng dụng và các hệ thống chưa đủ, do dữ liệu được chuyển đổi thiếu mô tả tường minh về ngữ nghĩa của nó. Sự tích hợp của các ứng dụng cũng phải bao gồm sự tích hợp cả về ngữ nghĩa.

HTTP và HTML đã cung cấp các cách để có thể nhận thông tin và trình diễn các tài liệu siêu văn bản. Tuy nhiên, có một khối lượng khổng lồ các tài nguyên thông tin trên Web, điều này làm nảy sinh vấn đề là làm thế nào để tìm kiếm chính xác tài nguyên mình mong muốn. Dữ liệu trong các file HTML có thể hữu ích ở ngữ cảnh này nhưng vô nghĩa đối với ngữ cảnh khác. Nhân loại đang dần dần tiến đến cuộc cách mạng công nghệ 4.0 và ngày các công nghệ thông tin và truyền thông đã có khả năng để thu thập được một số lượng lớn dữ liệu mà chúng có liên quan đến nhau về mặt khái niệm, tuy nhiên đa số những mối quan hệ này chỉ được con người "nhớ" chứ không được lưu trữ theo một cách mà giúp các máy tính có thể hiểu để xử lý. Thách thức này đã chỉ ra một hướng nghiên cứ đó là tạo ra khả năng

cho phép con người tạo, lưu giữ, sắp xếp, ghi phụ chú và truy xuất kho dữ liệu cá nhân rất lớn của mỗi người trong quá khứ theo hình thức như một nhật ký cuộc sống được cá thể hoá và sẽ trở thành một sự bổ sung và trợ giúp cho bộ nhớ con người.

Những hoạt động này đều đặc biệt không được hỗ trợ tốt của các công cụ phần mềm. Ngoài sự tồn tại của các liên kết để thiết lập các liên kết giữa các tài liệu, thì các công cụ có giá trị nhất trên Web hiện nay là các bộ tìm kiếm (search engines). Các công cụ tìm kiếm theo từ khoá như Yahoo! và Google là các công cụ chính trong việc sử dụng Web hiện nay. Rõ ràng rằng Web sẽ không thành công lớn như hiện này nếu không có các công cụ tìm kiếm. Tuy nhiên, vẫn tồn tại các vấn đề liên quan đến các ứng dụng của chúng:

- Truy hồi cao, độ chính xác thấp: Ngay cả khi các trang liên quan chính được truy xuất, thì chũng vẫn không hữu ích khi rất rất nhiều các tài liệu ít liên quan hoặc không liên quan cũng được lấy về. Quá nhiều cũng dẫn đến không tốt cũng như quá ít.
- Truy hồi thấp hoặc không có. Trường hợp này xảy ra chúng ta không có được câu trả lời từ yêu cầu của chúng ta, hoặc các tài liệu liên quan và quan trọng không được lấy về. Cho dù việc truy hồi thấp khá hiếm khi có đối với các công cụ tìm kiếm, nhưng nõ vẫn xảy ra.
- Các kết quả rất nhạy cảm với từ vựng. Thông thường các từ khoá tìm kiếm ban đầu không cho ta kết quả như mong muốn, lý do là các tài liệu liên quan sử dụng các thuật ngữ khác với truy vấn của chúng ta. Điều này rõ ràng là không thoả mãn bởi vì các truy vấn cùng ngữ nghĩa nên cho cùng một kết quả.
- Kết quả chỉ là những trang Web đơn giản. Nếu chúng ta cần những thông tin dàn trãi trong các tài liệu khác nhau, chúng ta phải thực hiện nhiều truy vấn khác nhau để tập hợp các tài liệu liên quan; sau đó chúng ta sẽ xử lý bằng tay để trích rút các thông tin từng phần rồi kết hợp chúng lại với nhau.

1.2 Web ngữ nghĩa

1.2.1 Khái niệm

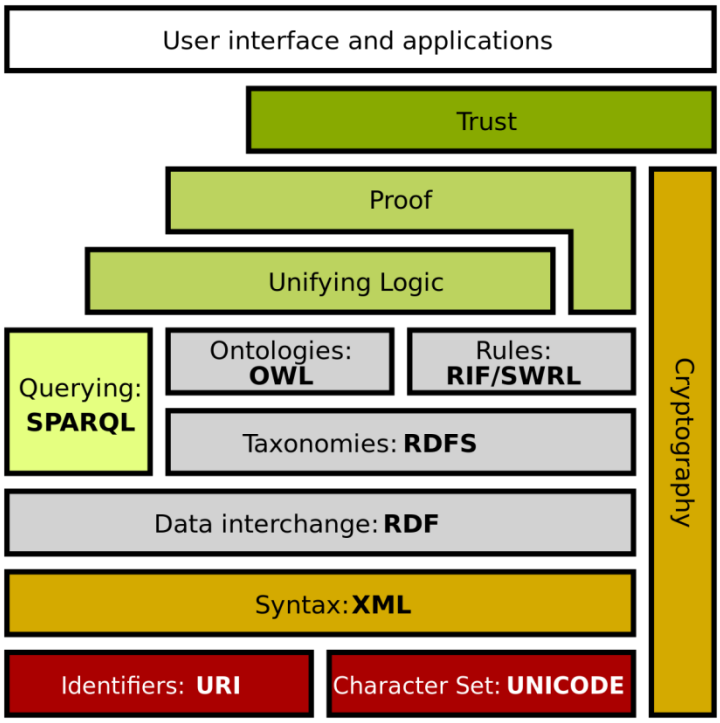
Web ngữ nghĩa không là Web riêng biệt mà là một sự mở rộng của Web hiện tại, theo cách thông tin được xác định ý nghĩa tốt hơn, nó cho phép máy tính và người cộng tác với nhau tốt hơn. Web ngữ nghĩa được hình thành từ ý tưởng của Tim Berners-Lee, người phát minh ra WWW, URI, HTTP, và HTML. Web ngữ nghĩa là một mạng lưới các thông tin được liên kết sao cho chúng có thể được xử lý dễ dàng bởi các máy tính ở phạm vi toàn cầu. Nó được xem là cách mô tả thông tin rất hiệu quả trên World Wide Web, và cũng được xem là một cơ sở dữ liệu có khả năng liên kết toàn cầu. Web ngữ nghĩa là một phương pháp cho phép định nghĩa và liên kết dữ liệu một cách có ngữ nghĩa hơn nhằm phục vụ cho máy tính có thể "hiểu" được. Web ngữ nghĩa còn cung cấp một môi trường chia sẻ và xử lý dữ liệu tự động bằng máy tính.

1.2.2 Siêu dữ liệu

Metadata (siêu dữ liệu) dùng để mô tả tài nguyên thông tin. Thuật ngữ "meta" xuất xứ là một từ Hy Lạp đùng để chỉ một cái gì đó có bản chất cơ bản hơn hoặc cao hơn. Một định nghĩa chung nhất và được dùng phổ biến trong cộng đồng những người làm Công nghệ Thông tin: "Metadata là dữ liệu về dữ liệu khác" (Metadata is data about other data) hay có thể nói ngắn gọn là dữ liệu về dữ liệu.

1.2.3 Kiến trúc Web ngữ nghĩa

Web ngữ nghĩa là một tập hợp/một chồng (stack) các ngôn ngữ. Tất cả các lớp của Web ngữ nghĩa được sử dụng để đảm bảo độ an toàn và giá trị thông tin trở nên tốt nhất.



Hình 1.2.3. **Kiến trúc Web ngữ nghĩa**

1.2.4 Các khái niệm cơ bản của Web ngữ nghĩa

1.2.4.1 Thực thể có tên

Thực thể có tên là con người, tổ chức, nơi chốn và những đối tượng khác được tham khảo đến bằng tên. Thực thể có tên khác về mặt bản chất lẫn ngữ nghĩa với các từ ở chỗ nó được dùng để chỉ các cá thể riêng biệt còn các từ được dùng để chỉ các khái niệm, quan hệ, thuộc tính nói chung

1.2.4.2 Tài nguyên

Thuật ngữ "tài nguyên" hay "resource" trên Web là một phạm trù rộng lớn dùng để chỉ mọi đối tượng có thể tìm thấy trên Web như khái niệm, từ vựng, thực thể, tính chất và quan hệ giữa các đối tượng.. Tài nguyên trên Web cũng chính là dữ liệu của trang Web đó, và là mục tiêu nghiên cứu của Web ngữ nghĩa.

Tài nguyên trên Web là khái niệm rộng hơn thực thể có tên. Và cũng như thực thể có tên, cùng một tài nguyên nhưng có thể được đặc tên khác nhau trong khi có nhiều tài nguyên bản chất khác nhau nhưng lại có cùng tên. Điều đó nảy sinh yêu cầu định danh mỗi tài nguyên bằng một định danh duy nhất. Các tài nguyên khác

nhau sẽ có định danh khác nhau. Định danh này được gọi là một URI (Uniform Resource Identifier).

1.2.4.3 Định danh tài nguyên

Tài nguyên được định danh bằng URI (Uniform Resource Identifier - định danh tài nguyên thống nhất). URI là một chuỗi các ký tự dùng để định danh tài nguyên trên Internet. Những dịnh danh này có khả năng tương tác với sự biểu diễn của tài nguyên trên mạng sử dụng giao thức cụ thể và phổ biến nhất là HTTP, do đó thường gọi là lược đồ HTTP URI. Có nhiều lược đồ khác ngoài HTTP URI như: ftp, tel, urn, mailto.

Một URI gồm có nhiều thành phần được liệt kê dưới đây:

- Lược đồ URI thường là tên giao thức (chẳng hạn http, ftp, news, mailto). Ở đây thường là lược đồ HTTP URI.
- Tên miền (ví dụ: http://www.portal.ptit.edu.vn).
- Chỉ định thêm cổng (có thể không cần nếu là cổng mặc định của dịch vụ).
- Đường dẫn tuyệt đối trên máy phục vụ của tài nguyên (ví dụ: thumuc/trang).
- Các truy vấn (tuỳ chọn).
- Chỉ định mục con (tuỳ chọn).

1.3 Ngôn ngữ Cơ cấu mô tả tài nguyên

1.3.1 Ngôn ngữ mô tả tài nguyên RDF

Ngôn ngữ Cơ cấu mô tả tài nguyên - RDF được đề xuất nhằm khắc phục những nhược điểm của XML không thể giải quyết được. Định nghĩa cơ bản của ngôn ngữ RDF là dùng để mã hóa các siêu dữ liệu của các tài nguyên vào một bộ ba (RDF Triple): [chủ ngữ], [vị ngữ] và [đối tượng]. Ta biết rằng mỗi một thực thể hay khái niệm đều có các thuộc tính, mỗi thuộc tính đều có các giá trị, vì vậy mọi tài nguyên cũng đều có thể được biểu diễn qua ngôn ngữ RDF.

RDF mô tả các nguồn tài nguyên bởi bộ ba [chủ ngữ], [vị ngữ], [đối tượng]. Một [vị ngữ] là một khía cạnh, tính chất, thuộc tính, hay mối liên hệ mô tả cho một tài nguyên. Một phát biểu bao gồm một tài nguyên riêng biệt, một thuộc tính được đặt tên, và giá trị thuộc tính cho tài nguyên đó ([đối tượng]). Giá trị này cơ bản có

thể là một tài nguyên khác hay một giá trị mang tính nghĩa đen hay dạng chuỗi văn bản tùy ý. [Chủ ngữ] và đối tượng được xác định qua Định danh tài nguyên thống nhất – URI, chẳng hạn chúng có thể là một liên kết của một trang web. Các [vị ngữ] cũng được xác định qua URI, do đó bất kì ai cũng có thể định nghĩa ra một khái niệm mới, một thuộc tính mới, bằng cách chỉ cần định nghĩa URI cho chúng. Bởi vì RDF sử dụng URI để biểu diễn các thông tin trong một tài liệu, các URI đảm bảo rằng các khái niệm không chỉ chứa văn bản thuần túy mà nó còn là định danh tài nguyên duy nhất mà tất cả người dùng có thể tìm kiếm được trên mạng. Trong RDF, các URI đóng một vai trò rất quan trọng: Chúng ta có thể tạo ra các (siêu) dữ liệu dựa trên bất kỳ một nguồn tài nguyên nào trên Web, ngữ nghĩa được đưa vào các nguồn tài nguyên Web thông qua các URI, và URI cho phép liên kết giữa các phần tử dữ liệu thông qua các thuộc tính.

1.3.1.1 Mô hình dữ liệu RDF

Mô hình cơ bản của RDF gồm ba bộ phận sau:

- Tài nguyên: là tất cả những gì được mô tả bằng biểu thức RDF
- Thuộc tính: là đặc tính hay quan hệ mô tả tính chất tài nguyên
- Phát_biểu: mỗi phát biểu gồm ba thành phần sau

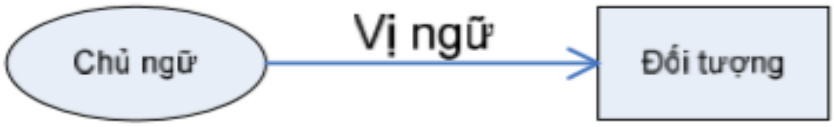
- [Chủ ngữ]: địa chỉ hay vị trí tài nguyên muốn mô tả.
- [Vị ngữ]: xác định tính chất của tài nguyên.
- [Đối tượng]: nội dung gán cho thuộc tính.

1.3.1.2 RDF và Cơ sở dữ liệu quan hệ

Trong các Cơ sở dữ liệu quan hệ truyền thống, dữ liệu được lưu dưới dạng các bảng. Trong mỗi bảng, mỗi hàng là một bản ghi không có giới hạn về số lượng các trường.

1.3.1.3 Đồ thị RDF

Tập hợp các bộ ba tạo thành đồ thị RDF (đồ thị có hướng). Các nút trong đồ thị là [chủ ngữ] và [đối tượng], các cung trong đồ thị là [Vị ngữ] và luôn có hường từ [chủ ngữ] tới [đối tượng]. Dùng đồ thị làm cho thông tin thể hiện rõ ràng và dễ hình dung hơn.



Hình 1.3.1.3. Đồ thị RDF

1.3.1.4 Cú pháp của RDF

Mô hình RDF thể hiện một mô hình ở mức trừu tượng để định nghĩa siêu dữ liệu. Cú pháp RDF được dùng để tạo ra và trao đổi siêu dữ liệu, vì thế RDF dựa trên cú pháp XMLDòng 7: cho biết kết thúc của thẻ rdf:RDF bắt đầu ở dòng 2 và cũng là thẻ kết thúc của tài liệu RDF

Để mô tả tập hợp của nhiều đối tượng như một bài báo khoa học được viết bởi nhiều tác giả, danh sách các sinh viên trong một khóa học, v.v... RDF cung cấp nhiều kiểu và nhiều thuộc tính tích hợp sẵn giúp mô tả được những tập như vậy, trong đó có kiểu khai báo “bộ chứa” (container), dùng để lưu danh sách các tài nguyên hoặc các kiểu giá trị (Một bộ chứa là một nguồn tài nguyên chứa những cái gì đó, những cái gì đó được đặt trong bộ chứa được gọi là các thành viên).Các phần tử của một bộ chứa có thể là các tài nguyên URI (có thể là rỗng) hay là các giá trị kiểu chuỗi kí tự. RDF định nghĩa 3 loại đối tượng “bộ chứa”: Bag, Sequence, và Alternative.

1.3.2 Lược đồ RDF vzà truy vấn RDF

1.3.2.1 Lược đồ RDF

RDF cung cấp một cách để mô tả các phát biểu đơn giản về các resource, sử dụng các thuộc tính và giá trị đã được định nghĩa trước. Tuy nhiên, RDF chỉ cho phép định nghĩa các quan hệ, chứ không nói rõ chỉ có thể có các loại quan hệ nào, hay các kiểu đối tượng có thể có trong miền hiện tại. Để làm được điều này, chúng ta sẽ phải sử dụng một phiên bản mở rộng của RDF, gọi là lược đồ RDF (RDF Scheme - RDFS). Lược đồ RDF là một ngôn ngữ bản thể luận dạng đơn nhất, nó cung cấp một khung để mô tả các lớp, thuộc tính của ứng dụng cụ thể. Các lớp trong RDFS giống như các lớp trong lập trình hướng đối tượng, cho phép các tài nguyên được định nghĩa như là một thực thể của lớp, hay lớp con của lớp.

Để thực hiện phân chia các lớp và các lớp con, RDFS sử dụng các phần tử như: rdfs:Class và rdfs:subClassOf.

1.3.2.2 Định nghĩa lớp

Các tài nguyên trên Web có thể chia thành các nhóm gọi là lớp. Các thành viên của nhóm được xem như là thể hiện của lớp đó. Thông qua các định danh URI, các tài nguyên được truy xuất và có thể được mô tả bằng các thuộc tinh RDF. Thuộc tính rdf: type được sử dụng để chỉ ra một tài nguyên là một thể hiện của một lớp.

1.3.2.3 Định nghĩa thuộc tính

Mô tả các tính chất của khái niệm. Lược đồ RDF cung cấp một bộ từ vựng để mô tả làm thế nào mà các thuộc tính và lớp có thể được sử dụng cùng nhau trong RDF

1.3.2.4 Ngôn ngữ truy vấn RDF

RDF là một cách để mô tả thông tin về các tài nguyên Web một cách linh động. Với lượng thông khổng lồ trên Web cần phải có ngôn ngữ truy vấn các tài liệu RDF một cách nhanh chóng và chính xác. Tổ chức W3C đã phát triển ngôn ngữ truy vấn trong các tài liệu RDF dựa trên cú pháp của ngôn ngữ truy vấn SQL trong CSDL quan hệ. Trong phần này sẽ giới thiệu sơ lược ngôn ngữ truy vấn RDF thông dụng là SPARQL. Chi tiết về ngôn ngữ SPARQL xem tại http://www.w3.org/TR/rdf-sparql-query/

SPARQL là một ngôn ngữ để truy cập thông tin từ các lược đồ RDF. Nó cung cấp các tính năng sau:

- Trích thông tin từ các dạng của URI
- Trích thông tin từ các lược con
- Xây dựng đồ thị RDF mới dựa trên thông tin trong đồ thị truy vấn

1.3 Linked Data

1.3.1 Khái niệm về Linked Data

Trong hoạt động tính toán máy tính, Linked data mô tả một phương thức tạo ra dữ liệu có cấu trúc để có thể liên kết được với nhau và trở nên có ích. Linked data được xây dựng dựa trên các tiêu chuẩn công nghệ Web như HTTP và URI để mở

rộng khả năng chia sẻ thông tin theo cách có thể được đọc tự động từ các máy tính hơn là việc chia sẻ nội dung trên các trang web để phục vụ cho người dùng. Điều này cho phép các nguồn tài nguyên khác nhau được kết nối và truy vấn.

1.3.2 Quy tắc Linked Data

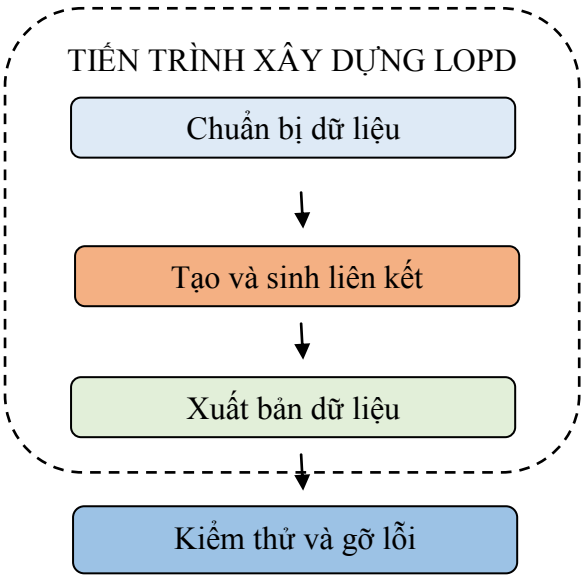
Tim Berners-Lee liệt kê bốn quy tắc triển khai Linked data trong các thảo luận của mình được tóm gọn trong các dòng sau:

1. Dùng URI để định danh mọi tài nguyên.
2. Sử dụng HTTP URI để các tài nguyên này được tham chiếu và tìm kiếm ("tham chiếu lại") bởi mọi người hay các ứng dụng.
3. Cung cấp thông tin hữu ích về các tài nguyên khi các URI của nó được tham chiếu lại, sử dụng các định dạng tiêu chuẩn như RDF/XML.
4. Bao hàm các liên kết đến liên kết khác, các URL có quan hệ bên trong dữ liệu tìm thấy để phát triển khả năng tìm kiếm các thông tin liên quan khác trên Web.

Lưu ý rằng mặc dù luật quy tắc thứ ba đề cập đến "các định dạng tiêu chuẩn", nó không cần bất cứ tiêu chuẩn đặc tả nào cả, chẳng hạn như RDF/XML.

Chương 2 - ỨNG DỤNG QUẢN LÝ VÀ TRUY XUẤT TÀI NGUYÊN THÔNG TIN TRONG TRƯỜNG ĐẠI HỌC– LINKED OPEN PTIT DATA (LOPD)

2.1 Tiến trình xây dựng ứng dụng LOPD



Hình 2.1. Tiến trình xuất bản Linked Data lên Web

Qua các khái niệm và cũng như các định nghĩa thì tôi đưa ra tiến trình để xây dựng ứng dụng LOPD gồm các tiến trình ở trên: "chuẩn bị dữ liệu, tạo và sinh liên kết, xuất bản dữ liệu, kiểm thử và gỡ lỗi". Các bước thực hiện trong quy trình ở hình 2.1 yêu cầu phải mật thiết và có sự tương tác mạnh mẽ với nhau để chính xác theo nguyên lý Linked Data.

2.2 Jena

2.2.1 Giới thiệu

Jena là một Java framework dùng cho việc xây dựng các ứng dụng web ngữ nghĩa. Cung cấp một môi trường lập trình cho RDF, RDFS and OWL, SPARQL. Bao gồm các công cụ suy diễn từ các luật cơ sở (rule-based inference engine). Open source Phiên bản đầu tiên ra đời 2000 do HP Lab xây dựng Jena 2 ra đời vào 2003 – phiên bản hiện tại là Jena 3.14.0

Jena Framework bao gồm:

- RDF API: Giao diện lập trình cho RDF
- OWL API: Giao diện lập trình cho OWL
- Xuất và đọc các file RDF theo dạng RDF/XML, N3 và N-Triples
- Cho phép lưu trữ trong bộ nhớ, lưu trữ cố định trong các tập tin đơn hay trong các hệ quản trị cơ sở dữ liệu (MySQL, HSQLDB, PostgreSQL, Oracle, Microsoft SQL Server)
- Hệ truy vấn SPARQL

2.2.2 Jena API

Kiến trúc của Jena đã được thiết kế để cho phép tích hợp một cách dễ dàng các thành phần xử lý có thể thay thế như bộ phân tích dữ liệu (parser), xuất bản (writer), lưu trữ và truy vấn.

Jena API bao gồm một tập các giao diện Java mô tả các nguồn tài nguyên (resources), các thuộc tính (properties), các phát biểu (statements) và các mô hình (model) dưới dạng khung mô tả tài nguyên RDF (Resource Description Framework)

2.2.3 Kiến trúc Jena 3

RDFNode interface: Cung cấp các thành phần của các bộ ba RDF {predicate, subject, object}.

Literal interface: Dùng để mô tả các chuỗi và chuyển đổi một số kiểu của Java như String, Int, và Double. Các đối tượng hiện thực giao diện Property có thể là một predicate trong bộ ba {predicate, subject, object}.

Statement interface: mô tả bộ ba {predicate, subject, object}. Đồng thời nó cũng có thể được dùng như một đối tượng .

Các đối tượng hiện thực các giao diện Container, Alt, Bag, hoặc Seq có thể là đối tượng .

2.3 Mô hình hóa thông tin với Jena

2.3.1 Chương trình Hello World! trong Jena

2.3.2 Tạo mô hình RDF

2.3.2.1 Tạo tài nguyên và thêm thuộc tính

2.3.3 Đọc mô hình RDF

2.4 Chuyển đổi dữ liệu web ngữ nghĩa

2.4.1 Dữ liệu từ Excel

2.4.2 Dữ liệu từ DBF

2.5 Chuẩn Dublin Core Metadata

Dublin Core Metadata3 là một chuẩn metadata được nhiều người biết đến và được dùng rộng rãi trong cộng đồng các nhà nghiên cứu, chuyên gia về thư viện số. Dublin Core Metadata lần đầu tiên được xuất năm 1995 bởi Dublin Core Metadata Element Initiative. Dublin là tên một địa danh Dublin, Ohio ở Mỹ nơi đã tổ chức hội thảo OCLC/NCSA Metadata Workshop năm 1995. Core có nghĩa là một danh sách các thành phần cốt lõi dùng mô tả tài nguyên (Element metadata), những thành phần này có thể mở rộng thêm.

Theo [12], tháng 9/2001 bộ yếu tố siêu dữ liệu Dublin Core Metadata được ban hành thành tiêu chuẩn Mỹ, gọi là tiêu chuẩn "The Dublin Core Metadata Element Set" ANSI/NISO Z39.85-2001.

Trong hệ thống của chúng tôi, những thông tin metadata sau được rút ra từ tài liệu:

- Creator (Author): Thông tin tên của các tác giả tài liệu.
- Title: tựa đề tài liệu.
- Description (Abstract): Tóm tắt nội dung của tài liệu.
- Publisher: Nơi công bố, xuất bản tài liệu.
- Source (DOI): Nơi download tài liệu hoặc địa chỉ chứa thông tin bài báo.
- Date (Year): Năm công bố, xuất bản tài liệu.

Chương 3 - PHÁT TRIỂN ỨNG DỤNG LOPD

3.1 Giới thiệu bài toán

3.1.1 Yêu cầu bài toán

Mục tiêu chính của ứng dụng là cải thiện kết quả tìm kiếm và truy xuất nhiều thông tin hơn liên quan cho mục đích sử dụng của người dùng. Đồng thời có thể trả lời được các câu hỏi như "Các bài báo trong năm 2018 của PGS TS Hoàng Hữu Hạnh?". Vì vậy, chương trình ứng dụng sẽ cung cấp và chia sẻ thông tin theo phương thức:

- Tìm kiếm thông tin: Hệ thống sẽ hỗ trợ tìm kiếm chính xác và gần đúng.

Với tìm kiếm chính xác, người dùng nhập đầy đủ và chính xác từ khóa cần tìm, lúc này hệ thống sẽ hiển thị tất cả lên các thông tin liên quan đến các giảng viên (Các điểm tương đồng, các bài báo cùng đề tài…).

Với tìm kiếm gần đúng, người dùng chỉ cần nhập một cụm từ liên quan đến các thực thể tồn tại trong hệ thống. Kết quả sẽ trả về tên các thực thể có liên quan và người dùng có thể chọn xem chi tiết để biết thông tin.

3.1.2 Phân tích vấn đề

Từ các hạn chế của các thư viện số hay các kho lưu trữ thông tin về các công trình, bài báo khoa học của các giảng viên cũng như các chủ đề tương đồng khi tìm kiếm gây cho việc tìm kiếm khó khăn và không tiếp cận được tối đa về thông tin tìm kiếm liên quan.

Hoặc như muốn sử dụng dữ liệu từ ác nguồn có sẵn để trả lời các câu hỏi như chủ đề này có những tác giả và bài báo nào tương đồng hay không. Hay các tác giả nào có các bài báo về chủ đề nào đó thì web bình thường khó có thể trả lời được câu hỏi này.

Từ những khó khăn trên tôi đưa ra giải pháp đó là sử dụng công nghệ web ngữ nghĩa để giải quyết bài toán. Công nghệ web ngữ nghĩa với đặc điểm lưu trữ dữ liệu dưới định dạng XML và mô hình dữ liệu thông minh nên việc lưu trữ dữ liệu có tính tùy biến cao và hỗ trợ tìm kiếm nhanh, thông tin chất lượng hơn.

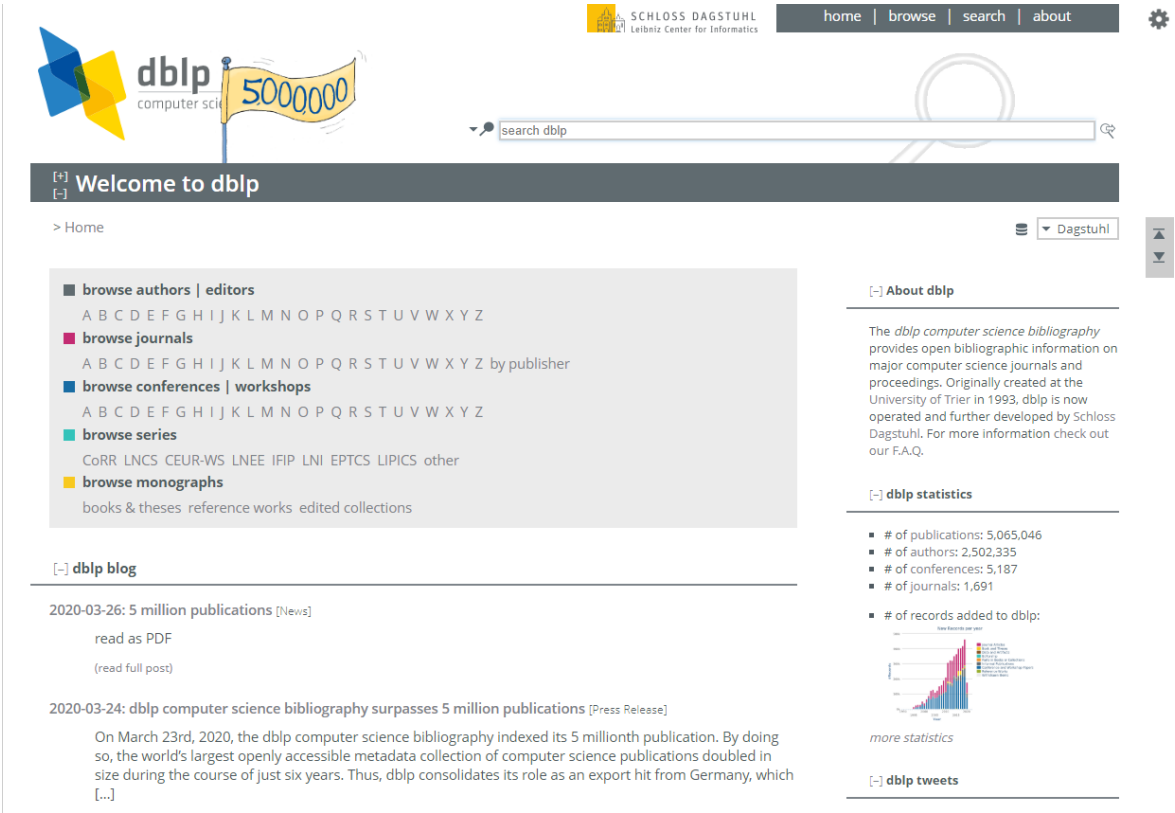
3.1.3 Chuẩn bị dữ liệu

3.1.3.1 Dữ liệu từ trường Đại học

Dữ liệu về tài nguyên thông tin trường Đại học bao gồm thông tin khoa học của các giảng viên, tác giả các bài báo, các công trình nghiên cứu khoa học các cấp. Dữ liệu từ file Excel gồm các thông tin các đề tài, giải thưởng, công trình nghiên cứu….

3.1.3.2 Dữ liệu từ DBLP Computer Science Bibliography

DBLP cung cấp thông tin về chỉ mục các bài báo trong lĩnh vực khoa học máy tính, hệ thống được phát triển bởi trường đại học Universität Trier của Đức. Tính đến tháng 1/2011 DBLP chứa thông tin chỉ mục của 1,5 triệu bài báo trong lĩnh vực khoa học máy tính được thu thập từ các thư viện số, các hội nghị và các tạp chí. Dữ liệu của DBLP được xuất ra các dạng CDF, XML và SQL, người phát triển có thể download các file này từ trên web của hệ thống.



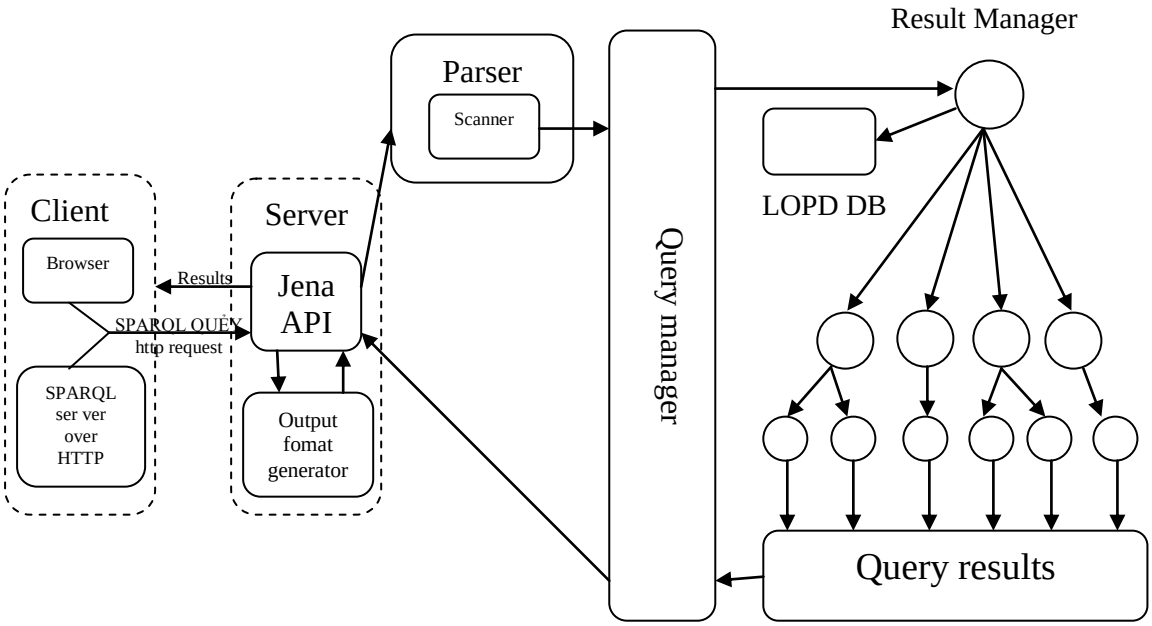
Hình 3.1.3.2. Cơ sở dữ liệu DBLP

3.2 Giải pháp hệ thống

3.2.1 Kiến trúc hệ thống

LOPD (Linked Open PTIT DATA): là một dịch vụ dùng SPARQL để truy vấn dữ liệu DBLP.

Hình 3.8 cho thấy tổng quan về kiến trúc của dịch vụ. Trong một vài trường hợp, Các hệ thống có thể dịch một truy vấn SPARQL tới một dãy cầu nối (lặp) các API, thu thập dữ liệu và trả lại cho người dùng. Dịch vụ hiện hỗ trợ SPARQL truy vấn thông qua HTTP GET. Bộ phân tích cú pháp sử dụng một máy quét để nhận dạng lexemes trong một truy vấn SPARQL và tạo ra các cấu trúc dữ liệu cần thiết bởi Query Manager. Mô-đun này có trách nhiệm tách các truy vấn vào các truy vấn con theo các điều khiển từ xa bởi các API sẵn có. ResultManager xử lý các truy vấn phụ và kết quả của nó để tạo ra bản đồ kết quả cuối cùng.



Hình 3.2.1. Quá trình thực hiện LOSM

3.2.2 Thiết kế cơ sở dữ liệu

3.2.2.1 Mô tả cấu trúc dữ liệu DBLP

Sau đây là cấu trúc bảng SQL của DBLP được Tiến sĩ Jörg Diederich xây dựng lên từ file XML của DBLP. Dữ liệu này được sử dụng trong hệ thống tìm kiếm Faceted DBLP và được cập nhật mỗi tuần một lần bằng cách sử dụng một đoạn script đọc dữ liệu trực tiếp từ file XML.

Dữ liệu được lưu trữ trong 3 bảng:

- dblp_pub_new: lưu thông tin bài báo.

Thông tin trong bảng bao gồm: tựa đề bài báo, năm xuất bản, số trang, tên nhà xuất bản, và một số thông tin định danh bài báo trong file XML của DBLP được giải thích chi tiết trong phần mô tả phía dưới. Trong cơ sở dữ liệu này mỗi bài báo có một id riêng được dùng chung cho các bảng có liên quan đến bài báo.

- dblp_author_ref_new: lưu thông tin về tác giả bài báo.

Thông tin trong bảng bao gồm: tên tác giả, tác giả có phải là một người biên tập (editor) hay không. Trong bảng này, những tác giả viết cùng một bài báo thì có id giống nhau và giống id tương ứng của bài báo cótrong bảng dblp_pub_new.

- dblp_ref_new: lưu thông tin về các tham chiếu (reference) giữa các bài báo.

Bài báo có id tương ứng trong bảng dblp_pub_new được tham chiếu bởi những bài báo nào được xác định bằng khóa dblp key.

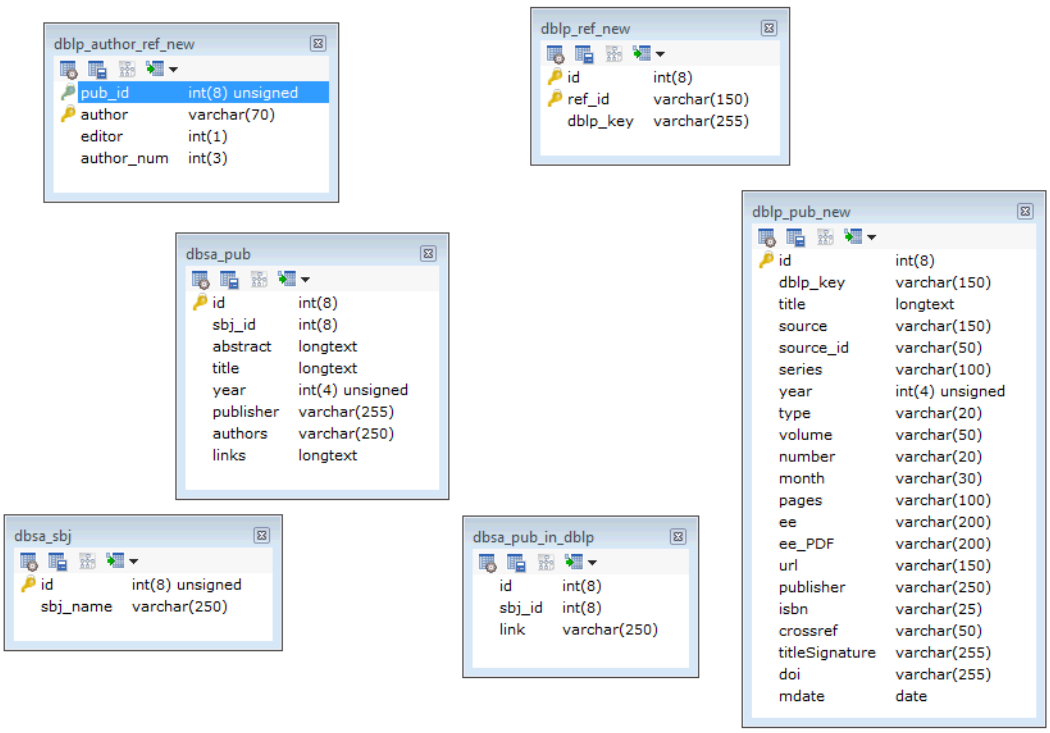
3.2.2.2 Cơ sở dữ liệu hệ thống

Như vậy trong cấu trúc bảng của dblp được trình bày ở trên, hệ thống không chứa dữ liệu về tóm tắt của bài báo (abstract).

Từ cấu trúc này nhóm bổ sung thêm vào cấu trúc những bảng sau để đảm bảo việc có thể cập nhật được dữ liệu mới của DBLP và có thể lưu được các thông tin về chủ đề, cũng như tóm tắt của bài báo

- dbsa_sbj**:** lưu thông tin về chủ đề của lĩnh vực khoa học máy tính.
- dbsa_pub: lưu thông tin bài báo được thu thập về từ các thư viện số.

- dbsa_pub_in_dblp: bảng lưu thông tin bổ sung của các bài báo trong dữ liệu dblp bao gồm chủ đề, những đường dẫn mở rộng (nơi mà bài báo có thể được tìm thấy – trang cá nhân của tác giả …).

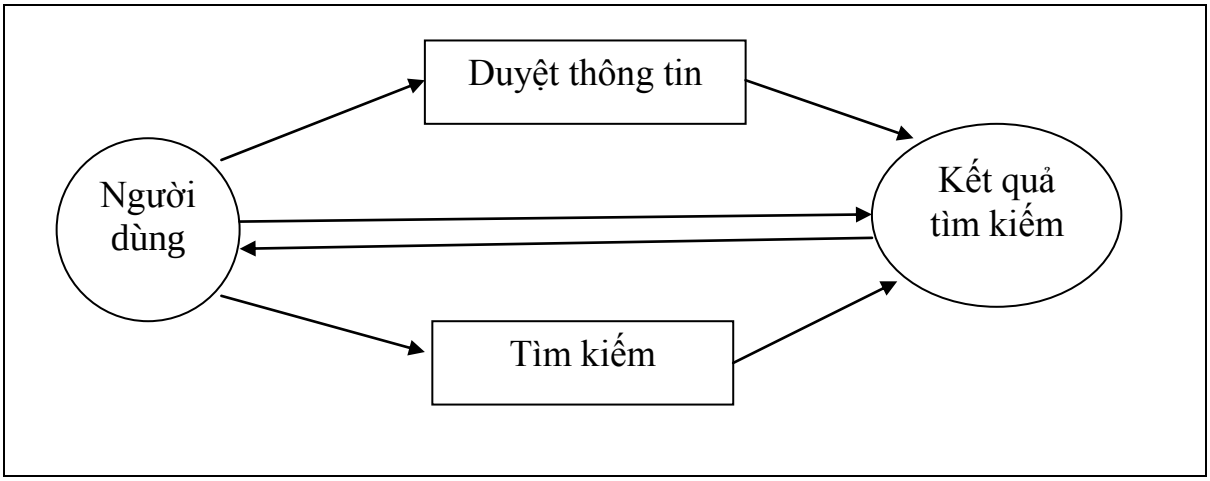


Hình 3.2.2.2. Các bảng trong cơ sở dữ liệu hệ thống

3.3 Xây dựng ứng dụng

3.3.1 Mô tả User case

Phần này sẽ xây dựng mô hình use-case nhằm cung cấp một cách chi tiết về các chức năng cơ bản như tìm kiếm thông tin, xem thông tin về một địa điểm cụ thể (duyệt thông tin) và xem các thông tin chi tiết của địa điểm. Mô hình use-case được thể hiện như hình sau:



Hình 3.3.1. Use case hệ thống tìm kiếm thông tin tác giả

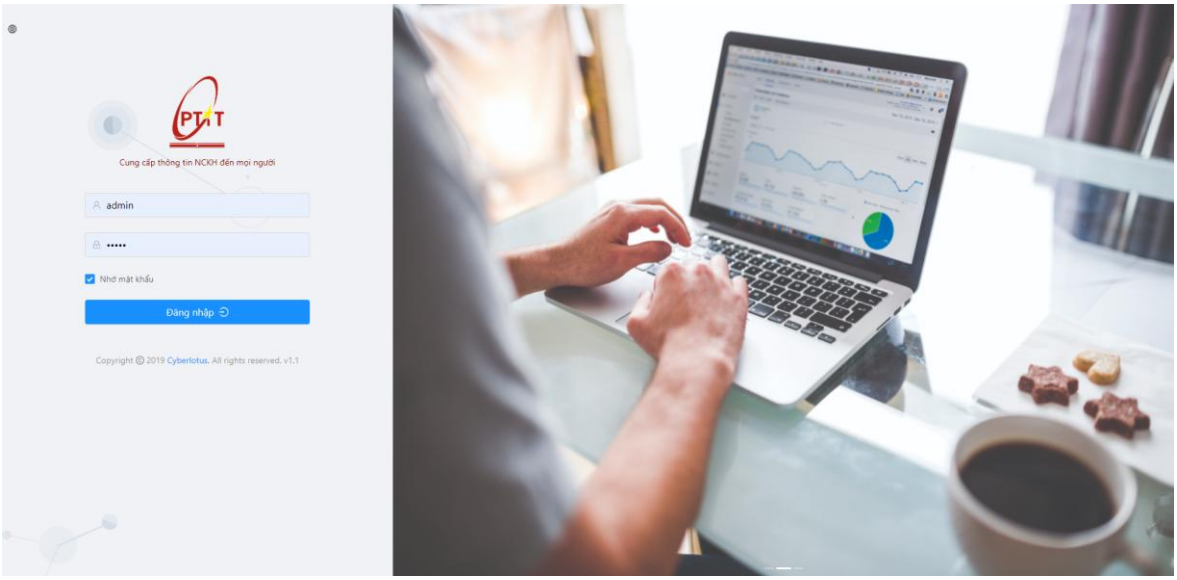
3.3.2 Đặc tả chức năng

- Chức năng tìm kiếm: Chức năng này cho phép người sử dụng tìm thông tin trong hệ thống bằng cách nhập từ khóa thông tin muốn tìm. Hệ thống duyệt file TACGIA.rdf tìm tất cả các tài nguyên liên quan đến từ khóa muốn tìm và trả về kết quả. Khi lấy được thông tin phù hợp sẽ hiển thị

- Chức năng duyệt thông tin: Chức năng này cho phép người sử dụng xem thông tin trong danh sách các tài nguyên liên quan đến từ khóa bằng cách chọn mục thông tin cần xem. Hệ thống tìm các thực thể liên quan đến mục thông tin được chọn và trả về các kết quả cho người sử dụng.

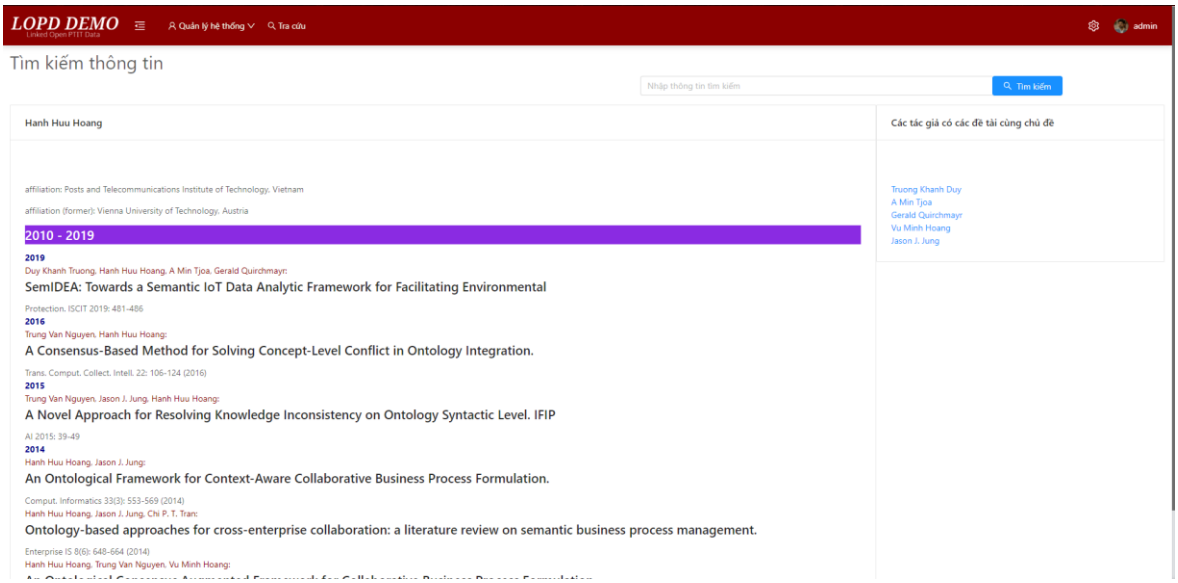
3.3.3 Thiết kế giao diện

3.3.3.1 Giao điện người dùng mặc định



Hình 3.3.3.1. Ứng dụng tìm kiếm thông tin tác giả

3.3.3.2 Kết quả tìm kiếm với tác giả



Hình 3.3.3.2. Kết quả tìm kiếm tác giả PGS.TS Hoàng Hữu Hạnh và các gợi ý các tác giả có các bài báo hay công trình cùng chủ đề

KẾT LUẬN VÀ KIẾN NGHỊ

Luận văn đã nghiên cứu và trình bày những kiến thức căn bản về web ngữ nghĩa như kiến thức về RDF, RDF Schema, ngôn ngữ truy vấn dữ liệu web ngữ nghĩa SPARQL, môi trường lập trình web ngữ nghĩa và trực quan hóa dữ liệu.

Qua đó luận văn đạt được một số kết quả như sau:

Về lý thuyết, luận văn đã đi sâu nghiên cứu được nhiều kiến thức về RDF và RDFS, từ đó hiểu được công nghệ web ngữ nghĩa để có thể dựa vào đó triển khai các ứng dụng khác. Trình bày cụ thể phương pháp lập trình với web ngữ nghĩa, cách chuyển đổi dữ liệu cho web ngữ nghĩa. Luận văn cũng đã trình bày cách trích xuất và xử lý dữ liệu từ DBLP Computer Science Bibliography

Về ứng dụng minh hoạ, với mục tiêu làm rõ thêm lý thuyết, luận văn ứng dụng xây dựng web ngữ nghĩa với các công cụ hỗ trợ như Intellij và Maven. Cụ thể là xây dựng được dữ liệu RDF cơ bản về các tác giả khoa học dựa trên dữ liệu thu thập từ DBLP trích xuất dữ liệu và khai thác các tính năng truy xuất trên một tài liệu có mô tả ngữ nghĩa nhằm chia sẻ tài nguyên thông tin về các bài báo, công trình khoa học và thực hiện tìm kiếm với những kết quả chính xác hơn, đồng thời cũng tận dụng được hết được các nguồn tài nguyên trong hệ thống.

Hướng phát triển:

Xây dựng ontology để hỗ trợ quá trình truy xuất và tìm kiếm thông tin hiệu quả hơn.

Phát triển ứng dụng với chức năng bổ sung và cập nhật thông tin trong ontology.

Sử dụng các công cụ lập trình di động và dữ liệu để tạo ứng dụng truy cập trên thiết bị di động.