1. TỔNG QUAN
   1. Tổng quan đề tài
      1. Đặt vấn đề

Ngày nay internet đã và đang là nguồn kiến thức vô tận mang lại nhiều lợi ích cho con người. Sự phát triển mạnh mẽ của nó kéo theo việc những kiến thức trong ngành công nghệ thông tin tăng lên nhanh chóng làm cho việc tra cứu kiến thức cần thiết trở nên khó khăn hơn. Với các công cụ tìm kiếm hiện nay như Google, Yahoo… chỉ giúp người dùng tìm được những tài liệu có chứa từ khóa. Từ đây người dùng phải tốn thời gian và công sức vào từng tài liệu để tìm được đúng thông tin mình cần mà có khi không tìm thấy hoặc tìm thấy thông tin sai lệch. Vấn đề đặt ra là làm sao để có được một công cụ tìm kiếm theo ngữ nghĩa, hiểu được và trả lời câu hỏi của người dùng bằng ngôn ngữ tự nhiên một cách thân thiện. Đặc biệt có thể tìm kiếm bằng tiếng Việt, nhu cầu mà hầu như rất ít công cụ hỗ trợ và kết quả còn hạn chế [1].

Dùng Ontology là một giải pháp biểu diễn tri thức và chia sẻ thông tin mà cả hệ thống và con người có thể hiểu được. Ontology chứa những đặc tả rõ ràng các khái niệm về một lĩnh vực và quan hệ giữa các khái niệm đó [2]. Nó được dụng trong trí tuệ nhân tạo, công nghệ Web ngữ nghĩa (Semantic Web), các hệ thống kỹ thuật, kỹ thuật phần mềm, sinh tin học và kiến trúc thông tin như là một hình thức biểu diễn tri thức về thế giới hoặc một số lĩnh vực cụ thể [11].

Nhận thấy những nhu cầu trên, nhóm thực hiện đề tài :

**“Xây dựng và làm giàu Ontology tiếng Việt chuyên ngành Công nghệ thông tin”.**

Đề tài nhằm xây dựng một Ontology là nền tảng cho những ứng dụng sau này như tìm kiếm thông tin tiếng Việt, hệ thống hỏi đáp tiếng Việt cho ngành công nghệ thông tin, hỗ trợ cho web ngữ nghĩa, giúp ghi chú (annotation) cho tài liệu công nghệ thông tin tiếng Việt. Ontology này có khả năng mở rộng cấu trúc và dữ liệu để phục vụ mục đích hỏi đáp của người dùng. Ngoài ra nhóm cũng sẽ xây dựng công cụ cho phép làm giàu Ontology từ internet.

* + 1. Khảo sát các nghiên cứu có liên quan
       1. Thế giới
          1. TheComputingOntology [3]

Được công bố năm 2005, ban đầu có tên là Ontology Project được xây dựng bởi một nhóm thuộc tổ chức ACM [4] nhằm biểu diễn kiến thức về máy tính và thông tin có quan hệ chặt chẽ theo quy tắc phục vụ cho việc nghiên cứu hay giảng dạy trong lĩnh vực liên quan tới tính toán và quản lý, xử lý thông tin. Công việc được hỗ trợ bởi tổ chức khoa học quốc gia Mỹ (*National Science Foundation*), ACM (Association for Computing Machinery), IEEE và Đại học Mở của Hà Lan (*Open University of the Netherlands*) [3].

Nguồn dữ liệu gồm: Tài liệu chương trình đào tạo của các trường đại học như: Lewis University, Villanova University…[3] và những thuật ngữ quan trọng trong các môn học về tính toán từ hệ thống phân lớp trên ACM (ACM Computing Classification System). Được xây dựng dùng công cụ soạn thảo ontology là Protégé, đến nay đã có 6 phiên bản của ComputingOntology trên trang web chính thức của nó.

*Nhận xét:* Xây dựng được một ontology về tính toán và thông tin hỗ trợ phát triển hệ thống tư vấn về chương trình đào tạo, phát triển những chương trình học mới, kiểm tra những chương trình học đã có, làm rõ ràng các mối quan hệ giữa những môn học với nhau, phát triển những chương trình gồm nhiều ngành học, đóng góp cho việc phân lớp trong nghiên cứu. Tuy nhiên, dữ liệu của ontology là tiếng Anh không thể phục vụ cho các nghiên cứu chuyên ngành công nghệ thông tin tiếng Việt.

* + - * 1. SwetoDBLP

Được xây dựng bởi nhóm tác giả từ khoa Khoa học máy tính của trường Đại học Georgia, Mỹ. SwetoDblp là một ontology có kích thước lớn tập trung vào dữ liệu thông tin của các bài báo về khoa học máy tính như: Tên, tác giả, nhà xuất bản… Dữ liệu chính của nó lấy từ cơ sở dữ liệu DBLP [5]. Ngoài ra, còn có 3 nguồn dữ liệu khác được dùng để tạo SwetoDblp là danh sách các trường đại học lấy từ Google được chỉnh sửa bằng tay lại cho phù hợp, danh sách các website của nhà xuất bản và danh sách các hội thảo được tạo bằng tay theo dữ liệu trong DBLP.

Dữ liệu của SwetoDblp được lưu trữ dùng định dạng RDF, sử dụng bộ từ vựng lược đồ (schema-vocabulary) có sẵn như FOAF []<http://xmlns.com/foaf/spec/>] và Dublin Core [<http://dublincore.org/>]. Việc tạo ra và cập nhật ontology được thực hiện dùng công cụ SAX-parser để chuyển dữ liệu dạng XML của DBLP sang RDF. Dữ liệu sẽ được cập nhật hàng tháng theo dữ liệu XML mới nhất từ DBLP và danh sách các trường đại học, nhà xuất bản và hội thảo.

SwetoDblp hiện đang được sử dụng để kiểm tra cho OptARQ, một cơ chế cho phép tối ưu hóa câu truy vấn vào ontology dùng SPARQL [6]. Ngoài ra, ontology này còn được dùng để tìm kiếm các bài báo và chuyên gia, phân biệt, tránh sự nhập nhằng giữa tên các nhà nghiên cứu trong danh sách mail của DBWorld [7].

*Nhận xét:* Ontology này như một thư viện điện tử với lượng thông tin lớn về các bài báo, không phục vụ cho việc tìm kiếm các khái niệm và thông tin trong ngành công nghệ thông tin.

* + - 1. Trong nước
         1. Ontology for Vietnamese Language – Open version [8]

Là một ontology tổng quát (universal ontology) được thực hiện bởi Nguyễn Tuấn Đăng, Võ Hoài An, Nguyễn Trí Phúc trường Đại học Công nghệ Thông tin. Xây dựng trên phiên bản Protégé 3.4.3. Mục tiêu tác giả xây dựng Ontology này là để đóng góp cho những nghiên cứu về xử lý ngôn ngữ tiếng Việt, xây dựng tri thức mở rộng trong nhiều lĩnh vực bằng tiếng Việt.

Dữ liệu của Ontology là dữ liệu tổng quát về các lĩnh vực gồm 10 lĩnh vực chính theo các mục được lấy theo VNExpress như: Khoa học, Pháp luật, Chính trị, Kinh doanh, Thể thao, Văn hóa du lịch, Xã hội, Vi tính, Viễn thông, Ô tô xe máy. Ngoài ra còn lấy dữ liệu từ các nguồn như Wikipedia tiếng Việt, Yellow Page và nhiều website khác nhau liên quan đến các lĩnh vực trên [9].

*Nhận xét:* Kết quả tạo ra được Ontology gồm số lượng lớp là 2.543, số lượng cá thể là 10.024, với 312 ràng buộc và 87 thuộc tính thuộc nhiều lĩnh vực. Tuy nhiên, dữ liệu của ontology mang tính phổ quát, không tập trung vào một lĩnh vực (domain) cụ thể. Ví dụ như trong ngành Công nghệ thông tin không có chứa thông tin về những khái niệm, chuyên gia hay chương trình đào tạo của ngành.

* + - * 1. Ontology khoa học công nghệ [10]

Được thực hiện bởi Bộ môn Hệ thống thông tin của trường Đại học Bách khoa Hà Nội. Hệ thống hỗ trợ tìm kiếm dựa trên từ khóa, cấu trúc dữ liệu lưu trữ, tìm kiếm mở rộng dựa trên ngữ nghĩa và tri thức phục vụ cho việc quản lý tài liệu và thông tin trong lĩnh vực khoa học công nghệ. Nhằm giải quyết cho những yêu cầu đó tác giả đã đề xuất phương pháp xây dựng một ontology chuyên ngành khoa học công nghệ để khai thác các suy diễn ngữ nghĩa.

Những khái niệm được xây dựng dựa trên việc khảo sát nhu cầu quản lý thông tin tại phòng KHCN thuộc Đại học Bách Khoa Hà Nội, phòng KHCN thuộc sở Khoa học Công nghệ Thành Phố Hà Nội, sở Bưu chính Viễn thông. Người bảo trì có thể là tác giả hoặc những người có quan tâm và có kiến thức về Ontology sẽ nâng cấp cập nhật thông tin khi có thay đổi.

Với việc sử dụng Ontology này hệ thống ngoài việc dùng để tra cứu các đề tài, sản phẩm công nghệ, chuyên gia, tài liệu, giải pháp, công nghệ… thì còn có thể trả lời được những câu hỏi tổng hơp phân tích 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 như chuyên gia vừa trong lĩnh vực CNTT vừa trong lĩnh vực Hoá sinh.

Ontology này được xây dựng dùng phần mềm soạn thảo cơ sở tri thức được viết dựa trên các API của Protégé. Cơ sở dữ liệu này chứa dữ liệu về khoảng 3000 chuyên gia, 1500 đề tài cùng với hơn 150 lĩnh vực KHCN.

*Nhận xét:* Không rút trích được khái niệm hay cá thể từ nội dung tài liệu hay bài báo khoa học. Dữ liệu của Ontology KHCN không đi sâu vào chuyên ngành Công nghệ thông tin.

* + 1. Mục tiêu đề tài
* Xây dựng Ontology chuyên ngành công nghệ thông tin tiếng Việt phục vụ cho việc:
* Tìm kiếm thực thể có tên, xác định quan hệ giữa các thực thể
* Giúp hiểu được câu hỏi bằng ngôn ngữ tự nhiên và đưa ra câu truy vấn phù hợp.
* Hỗ trợ trả lời cho hệ thống hỏi đáp thông minh.

Ví dụ:

Java là gì?

Trường nào đã đoạt giải vô địch cuộc thi Robocon năm 2009?

* Sử dụng cho web ngữ nghĩa.
* Sử dụng cho hệ thống tư vấn về chương trình đào tạo công nghệ thông tin.
* Tạo được công cụ làm giàu Ontology tự động từ nhiều nguồn trên internet, sách, ebook chuyên ngành và từ các ontologies có sẵn.
* ***Phạm vi đề tài:***

## Xây dựng Ontology tiếng Việt giới hạn trong lĩnh vực Công nghệ thông tin –Ontology for Vietnamese Information Technology (OVIT) nhằm lưu trữ:

## Các khái niệm trong lĩnh vực Công nghệ thông tin và quan hệ giữa chúng.

## Thông tin các công ty, trường học, tổ chức, hiệp hội, chuyên gia, các sự kiện trong ngành và quan hệ ngữ nghĩa giữa chúng.

## Các chương trình đào tạo Công nghệ thông tin tại các trường.

## Nguồn dữ liệu: từ ComputingOntology của nhóm nghiên cứu thuộc ACM, trang Wikipedia tiếng Việt, website Bộ thông tin và truyền thông, một số website báo điện tử, các bài báo lĩnh vực công nghệ thông tin tiếng Việt, website các trường học, tài liệu từ internet tìm được từ công cụ tìm kiếm như Google, Yahoo.

* + 1. Phương pháp nghiên cứu
    2. Kết quả dự kiến

Kiến thức: Nắm được khái niệm, cấu trúc, mục đích, ứng dụng, cách xây dựng một ontology. Các công cụ hỗ trợ xây dựng ontology hiện nay và sử dụng ngôn ngữ Java để xây dựng công cụ làm giàu ontology (OVIT) bán tự động.

Phương pháp: Xây dựng ontology dùng công cụ Protégé, sử dụng API của Google và Yahoo để tìm kiếm nguồn tài liệu làm giàu, dùng thuật toán SVM để phân loại tài liệu công nghệ thông tin tiếng Việt, rút trích các cá thể từ tài liệu đã phân loại và lưu vào ontology.

Dữ liệu: Dự kiến nhập bằng tay được khoảng 1000 lớp, 100 quan hệ và 100 cá thể, làm giàu cá thể bán tự động được 1000 cá thể.

Công cụ: Xây dựng được công cụ làm giàu cá thể bán tự động: rút trích cá thể từ tài liệu công nghệ thông tin tiếng Việt, cho phép người dùng kiểm tra, chỉnh sửa sau đó lưu trữ vào cơ sở dữ liệu theo cấu trúc của ontology.

* 1. Tổng quan về Ontology
     1. Định nghĩa về ontology

Theo wikipedia, từ “ontology” tạm dịch là “bản thể luận” có nguồn gốc từ triết học và được dùng trong nhiều lĩnh vực như triết học, khoa học máy tính, hệ thống kỹ thuật, kỹ thuật phần mềm, tin sinh học, khoa học thư viện, kiến trúc thông tin và các website ngữ nghĩa (Semantic web). Hiện nay, có một số định nghĩa về ontology được biết đến rộng rãi như:

* Theo quan điểm triết học, ontology là nghiên cứu về bản chất của sinh vật, sự tồn tại hoặc những sự vật thực tế, cũng như các loại sinh vật cơ bản và các mối quan hệ của chúng (wikipedia).
* Trong khoa học máy tính, một ontology là một đặc tả rõ ràng của một sự trừu tượng hóa (An ontology is an explicit specification of a conceptualization) [20].
* Một ontology định nghĩa một tập từ vựng cho những nhà nghiên cứu sử dụng khi cần chia sẻ thông tin trong một lĩnh vực. Nó bao gồm những định nghĩa của các khái niệm cơ bản trong một lĩnh vực và mối quan hệ giữa chúng mà máy có thể hiểu được [1].

Tóm lại, ontology gồm những khái niệm về một lĩnh vực cụ thể và các mối quan hệ giữa chúng. Một ontology về một lĩnh vực sẽ mô tả rõ ràng những thực thể giúp con người và máy có thể hiểu và suy luận được theo ngữ nghĩa trong phạm vi lĩnh vực đó.

* + 1. Vì sao phải xây dựng ontology?

Để chia sẻ kiến thức chung về kiến trúc thông tin giữa con người hoặc những tác tử phần mềm với nhau [20]. Nếu các hệ thống cùng chia sẻ chúng một ontology bên dưới thì dữ liệu do con người nhập vào tại hệ thống này sau khi được xử lý thông qua ontology có thể được tổng hợp, phân tích tại một hệ thống khác và cung cấp thông tin cho người khác.

Cho phép tái sử dụng kiến thức về một lĩnh vực. Sau khi xây dựng một ontology cho một lĩnh vực, những người khác có thể tái sử dụng và mở rộng, làm giàu thêm cho nó. Hoặc cũng có thể tích hợp những ontology có sẵn để mô tả nhiều phần của một ontology về một lĩnh vực lớn.

Làm rõ ràng những giả định thuộc chuyên ngành. Việc sử dụng một ontology ở bên dưới thay vì trong code sẽ giúp dễ dàng thay đổi những giả định thuộc chuyên ngành khi kiến thức về lĩnh vực này của chúng ta thay đổi. Nếu những giả định này được viết trong code thì sẽ gây khó hiểu và khó thay đổi, sửa chữa nhất là đối với những người không phải là chuyên gia lập trình.

Có thể phân tích và suy luận kiến thức chuyên ngành vì những thuật ngữ, khái niệm cũng như các mối quan hệ giữa chúng đề được khai báo, đặc tả trong ontology với cấu trúc có thể suy luận được theo ngữ nghĩa.

* + 1. Thành phần của ontology
* Các lớp (Classes) - Khái niệm  
  Các lớp là các nhóm, tập hợp các đối tượng trừu tượng. Chúng có thể chứa các cá thể, các lớp khác, hay là sự phối hợp của cả hai.  
  Các ontology biến đổi tuỳ thuộc vào cấu trúc và nội dung của nó: Một lớp có thể chứa các lớp con, có thể là một lớp tổng quan (chứa tất cả mọi thứ), có thể là lớp chỉ chứa những cá thể riêng lẻ, Một lớp có thể xếp gộp vào hoặc bị xếp gộp vào bởi các lớp khác. Mối quan hệ xếp gộp này được sử dụng để tạo ra một cấu trúc có thứ bậc các lớp, thường là với một lớp thông dụng nhất kiểu Thing ở trên đỉnh và các lớp rất rõ ràng kiểu 2002, Ford ở phía dưới cùng.
* **Các cá thể (Individuals)**

Là những đối tượng đại diện trong domain.Có một sự khác biệt quan trọng giữa Protégé và OWL là OWL không sử dụng Unique Name Assumption (UNA). Điều này có nghĩa là trong OWL hai tên khác nhau có thể tham chiếu tới cùng một cá thể. Ví dụ: “nữ hoàng Elizabeth” và “Elizabeth Windsor” là 2 tên khác nhau nhưng cùng chỉ một cá thể. Những cá thể còn có thể được coi như là những trường hợp của lớp.

* Các thuộc tính (Properties)  
  Các đối tượng trong ontology có thể được mô tả thông qua việc khai báo các thuộc tính của chúng. Mỗi một thuộc tính đều có tên và giá trị của thuộc tính đó. Các thuộc tính được sử dụng để lưu trữ các thông tin mà đối tượng có thể có. Ví dụ, đối với một cá nhân có thể có các thuộc tính: Họ\_tên, ngày\_sinh, quê\_quán, số\_cmnd…  
  Giá trị của một thuộc tính có thể có các kiểu dữ liệu phức tạp.
* Các mối quan hệ (Relations)  
  Một trong những ứng dụng quan trọng của việc sử dụng các thuộc tính là để mô tả mối liên hệ giữa các đối tượng trong ontology. Một mối quan hệ là một thuộc tính có giá trị là một đối tượng nào đó trong ontology.  
  Một kiểu quan hệ quan trọng là kiểu quan hệ xếp gộp (subsumption). Kiểu quan hệ này mô tả các đối tượng nào là các thành viên của các lớp nào của các đối tượng.  
  Hiện tại, việc kết hợp các ontology là một tiến trình được làm phần lớn là thủ công, do vậy rất tốn thời gian và đắt đỏ. Việc sử dụng các ontology là cơ sở để cung cấp một định nghĩa thông dụng của các thuật ngữ cốt lõi có thể làm cho tiến trình này trở nên dễ quản lý hơn. Hiện đang có các nghiên cứu dựa trên các kỹ thuật sản sinh để nối kết các ontology, tuy nhiên lĩnh vực này mới chỉ hiện hữu về mặt lý thuyết
  + 1. Làm thế nào để xây dựng một ontology?
  1. Tổng kết chương

1. CƠ SỞ LÝ THUYẾT

Tài liệu tham khảo:

[1] **Natalya F. Noy and Deborah L. McGuinness.** *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford University, Stanford, CA, 94305

[2] **Phạm Thị Mỹ Phượng, Từ Thị Ngọc Thanh.** *Tìm kiếm ngữ nghĩa ứng dụng trên lĩnh vực eDoc.*

[3] **Lê Thành Nhân, Võ Trung Hùng, Cao Xuân Tuấn, Hoàng Thị Mỹ Lệ.** *MATHIS – Hệ thống hỗ trợ tạo chú thích và tìm kiếm tài liệu khoa học*. Tạp chí khoa học và công nghệ, Đại học Đà Nẵng - Số 4(39).2010

[4**] Trần Đình Khang, Vũ Tuyết Trinh, Đỗ Đức Thành, Đỗ Thị Ngọc Quỳnh**. *Một phương pháp tìm kiếm dựa trên Ontology phục vụ quản lý thông tin khoa học công nghệ.* Bộ môn Hệ thống Thông tin, Trường Đại Học Bách Khoa Hà Nội.

[5] **Lương Quý Tịnh Hà.** *Xây dựng công cụ tìm kiếm tài liệu học tập bằng các truy vấn ngôn ngữ tự nhiên trên kho học liệu mở tiếng Việt.*

[6]*Tài liệu hướng dẫn phiên bản mã nguồn mở OVL – Open 1.0*

[7] **Nguyen Phi Minh Tri, Nguyen Tuan Dang.** *Building a Universal Ontology for Vietnamese Language.* Faculty of Computer Science, University of Information Technology.

[8] **Sean Bechhofer, Ian Horrocks, Daniele Turi.** *The OWL Instance Store: System Description*. Information Management Group, School of Computer Science,The University of Manchester

[9] **Bijan Parsia and Evren Sirin.** *Pellet: An OWL DL Reasoner.* MINDSWAP Research Group, University of Maryland, College Park, MD

[10] <http://www.acm.org/>

[11] <http://what.csc.villanova.edu/twiki/bin/view/Main/OntologyProject>

[12] <http://what.csc.villanova.edu/twiki/bin/view/Main/IntroducingGroupMembers>

[13] <http://www.acm.org/education/curricula-recommendations> [2001 -- 2005 curriculum recommendations]

[14] <http://what.csc.villanova.edu/twiki/bin/view/Main/SourcesUsed>

[15] <http://what.csc.villanova.edu/twiki/bin/view/Main/OWLFileInformation>

[16] <http://dblp.uni-trier.de/>

[17] <http://xmlns.com/foaf/spec/>

[18] <http://dublincore.org/>

[19] <http://knoesis.wright.edu/library/ontologies/swetodblp/>

[20] **Thomas R.Gruber.** *Toward Principles for the Design of Ontologies Used for Knowledge Sharing.* Stanford Knowledge Systems Laboratory, 701 Welch Road, Building C Palo Alto, CA 94304, gruber@ksl.stanford.edu.