ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Vũ Ngọc Trình

NGHIÊN CỨU TÍCH HỢP MÔ HÌNH DỮ LIỆU TRONG TRUNG TÂM DỮ LIỆU NGÀNH DẦU KHÍ

Chuyên ngành: Hệ thống Thông tin
Mã số: 62 48 01 04

TÓM TẮT LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội – 2017

Công trình được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Người hướng dẫn khoa học: .....

1. PGS.TS. Hà Quang Thụy, ĐH Công nghệ

2. PGS.TSKH. Nguyễn Hùng Sơn, ĐH Varsava, Ba Lan

Phản biện:.....

.....

Phản biện:.....

.....

Phản biện:.....

.....

Luận án sẽ được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại .....

vào hồi giờ ngày tháng năm

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nôi

MỞ ĐẦU

Tính cấp thiết của luận án

Dữ liệu ngày nay đang dần được coi như một nguồn tài nguyên thực sự, đóng vai trò nguồn nhiên liệu chủ chốt tương tự như "dầu mỏ của Thế kỷ 20" và dữ liệu đang tạo ra một nền kinh tế mới[1]. Tích hợp dữ liệu (*data integration*) có thể giúp doanh nghiệp chuyển đổi dữ liệu thành tài nguyên tạo doanh thu thực sự cho doanh nghiệp. Tích hợp ontology[2] là một thành phần quan trọng trong tích hợp dữ liệu. Tích hợp ontology được coi là một dạng tích hợp dữ liệu được tiến hành trên kiểu dữ liệu đặc biệt, đồng thời, tích hợp dữ liệu dựa trên ontology được nhận diện là một kỹ thuật tích hợp dữ liệu khá phổ biến. Tích hợp dữ liệu và tích hợp ontology luôn là các chủ đề khoa học và công nghệ nhận được sự quan tâm của cộng đồng nghiên cứu-triển khai trên thế giới, tạo động lực nghiên cứu và triển khai về tích hợp dữ liệu và tích hợp ontology. Tích hợp dữ liệu là chủ đề nghiên cứu của một số luận án Tiến sỹ trên thế giới, chẳng hạn như [Doan02, Aleksovski08, Dragisic17]. Luận án của Đoàn An Hải [Doan02], một trong năm luận án Tiến sỹ được nhận giải thưởng luận án Tiến sỹ xuất sắc của Hiệp hội máy tính ACM, cung cấp các phân tích sâu sắc về tiếp cận tích hợp mô hình dữ liệu, tập trung vào miền ứng dụng bất động sản. Các luận án [Aleksovski08, Dragisic17] định hướng tới các kỹ thuật tích hợp dữ liệu dựa trên ontology, theo đó tích hợp mô hình dữ liệu được tiến hành thông qua mối quan hệ giữa ontology của dữ liệu đích với ontology từ các nguồn dữ liệu. Tích hợp lược đồ dữ liệu và ứng dụng là một chủ đề nghiên cứu và triển khai còn mới mẻ ở Việt Nam. Hiện nay, chưa có Ontology dầu khí Tiếng Việt, nhưng có một số ontology trong các lĩnh vực khác như VN-KIM [TrucVien07], [Tru07], BioCaster [Collier10]. Theo khảo sát của P. A. Bernstein và cộng sự [Bernstein11], sự hội tụ các phướng pháp tích hợp lược đồ dữ liệu và tích hợp thể hiện dữ liệu, hầu hết các phương pháp tích hợp lược đồ dữ liệu đều bao gồm thao tác tích hợp dữ liệu mức thể hiện. Hơn nữa,

---

[1] https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy

[2] Ontology được một số học giả dịch sang tiếng Việt là "bản thể", tuy nhiên, do từ "bản thể" không gợi nghĩa hơn từ "ontology" cho nên luận án sử dụng từ nguyên gốc "ontology".

tích hợp ontology cung cấp một khung nhìn điển hình về tích hợp dữ liệu mức lược đồ. Căn cứ vào xu hướng nghiên cứu này, luận án "Nghiên cứu tích hợp mô hình dữ liệu trong trung tâm dữ liệu ngành dầu khí Việt Nam" tập trung vào bài toán tích hợp ontology và ứng dụng xây dựng một ontology dầu khí Anh-Việt tại Tập đoàn dầu khí quốc gia Việt Nam.

Nghiên cứu của luận án hướng tới **một số mục tiêu** sau đây. Thứ nhất, luận án cung cấp một khảo sát khái quát về các kỹ thuật tích hợp dữ liệu mức lược đồ và tích hợp ontplogy. Thứ hai, luận án đề xuất một số kỹ thuật tích hợp ontology dựa trên việc sử dụng các độ đo (điển hình là độ đo khoảng cách Google) và áp dụng các phương pháp học máy (điển hình là phương pháp học máy với chỉ ví dụ dương). Về cơ bản, các kỹ thuật được đề xuất đều hướng tới ứng dụng vào miền dữ liệu để kiểm chứng tính khả thi và hiệu quả của đề xuất. Cuối cùng, luận án xây dựng phần mềm Ontology Dầu khí ANH - VIỆT nhằm phục vụ công tác nghiệp vụ tại Viện Dầu khí Việt Nam.

Đối tượng nghiên cứu của luận án là các kỹ thuật tích hợp ontology nhằm đề xuất một số kỹ thuật mới tích hợp ontology cho miền dữ liệu dầu khí và xây dựng một ontology dầu khí Anh – Việt.

Phạm vi nghiên cứu của luận án được giới hạn ở phương pháp tích hợp ontology tập trung vào miền dữ liệu dầu khí.

Phương pháp nghiên cứu của luận án là nghiên cứu lý thuyết đề xuất các kỹ thuật tích hợp ontology, nghiên cứu thực nghiệm để kiểm chứng đánh giá các kỹ thuật được đề xuất và công bố các kết quả nghiên cứu trên các ấn phẩm khoa học có uy tín. Luận án tiến hành các nghiên cứu ứng dụng để xây dựng một ontology dầu khí Anh – Việt tại Viện dầu khí Việt Nam.

Đóng góp của luận án. Luận án tham gia vào dòng nghiên cứu về tích hợp dữ liệu trên thế giới và đạt được một số đóng góp bước đầu, tập trung vào các nghiên cứu về tích hợp ontology trong miền dữ liệu dầu khí. *Về phương diện lý thuyết*, luận án đề nghị ba kỹ thuật tích hợp ontology. Thứ nhất, luận án đề xuất hai phương pháp tích hợp dữ liệu là tích hợp dữ liệu dựa trên độ đo Google [VNTrinh2, VNTrinh4]. Thứ hai, trên cơ sở ứng dụng các thuật toán học máy (đặc biệt là kỹ thuật học máy với chỉ dữ liệu dương) [VNTrinh4, VNTrinh5], luận án đã đề xuất một thuật toán kết hợp độ đo Google và độ đo khoảng cách Cosine với thuật toán học máy với chỉ dữ liệu dương để tích hợp dữ

liệu, nâng cao hiệu quả của thuật toán. Thứ ba, luận án đề nghị một kỹ thuật tích hợp ontology dựa trên thuật toán học máy Maximum Entropy và Beam Search sử dụng các kho ngữ liệu chuẩn (corpus)[VNTrinh1]. *Về phương diện ứng dụng*, các kết quả nghiên cứu của luận án có đóng góp trực tiếp vào hệ thống tích hợp dữ liệu tại Viện Dầu khí Việt Nam. Một ontology Dầu khí ANH-VIỆT được xây dựng dựa trên việc tích hợp từ điển Anh -Việt với Wordnet Tiếng Anh và Wikipedia Tiếng Việt được sử dụng cho việc tra cứu, nghiên cứu, đào tạo trong hiện tại và là cơ sở cho việc mở rộng, tích hợp với các hệ thống dữ liệu khác (ví dụ hệ thống chia sẻ tri thức đang có tại Viện Dầu khí Việt Nam...) và các ontology dầu khí khác trên thế giới, trong tương lai. Luận án cũng cung cấp một nghiên cứu tổng quan về tích hợp lược đồ dữ liệu (nói chung) và tích hợp ontology (nói riêng).

Bố cục của luận án gồm phần mở đầu và năm chương nội dung, phần kết luận và danh mục các tài liệu tham khảo.

Chương 1 của luận án cung cấp một nghiên cứu khái quát về các kỹ thuật tích hợp lược đồ dữ liệu, ontology và tích hợp ontology, và tính cấp thiết của việc xây dựng Ontology dầu khí Anh-Việt tại Viện dầu khí Việt Nam. Chương 2 của luận án trình bày chi tiết cách tiếp cận của luận án về việc sử dụng độ đo khoảng cách Google vào tích hợp ontology. Việc ứng dụng mô hình đề xuất vào miền dữ liệu dầu khí cũng được giới thiệu. Trong Chương 3, luận án trình bày về một mô hình tích hợp ontology từ tài nguyên kho ngữ liệu sử dụng học máy Maximum Entropy và Beam search. Chương 4 của luận án trình bày chi tiết một mô hình tích hợp ontology dựa trên việc sử dụng các kỹ thuật học máy với dữ liệu dương và dữ liệu chưa gán nhãn. Chương 5 của luận án trình bày một mô hình năm bước để xây dựng hệ thống ontology dầu khí ANH - VIỆT. Mô hình này được xây dựng dựa trên các kết quả nghiên cứu của luận án.

CHƯƠNG 1. GIỚI THIỆU CHUNG VỀ TÍCH HỢP DỮ LIỆU VÀ TÍCH HỢP ONTOLOGY

1.1. GIỚI THIỆU CHUNG VỀ TÍCH HỢP DỮ LIỆU

1.1.1. Khái niệm tích hợp dữ liệu

Như đã được giới thiệu, theo Đoàn An Hải và cộng sự [Doan12], *tích hợp dữ liệu được xem là một tập các kỹ thuật cho phép xây dựng các hệ thống được khớp nối lại nhằm chia sẻ và tích hợp linh hoạt dữ liệu từ nhiều nguồn dữ liệu tự trị*. Mục tiêu của một hệ thống tích hợp dữ liệu là cung cấp sự truy cập thống nhất vào một tập các

nguồn dữ liệu tự trị và không đồng nhất.

1.1.2. Kỹ thuật tích hợp lược đồ dữ liệu

Mỗi hướng tiếp cận tích hợp mô hình dữ liệu luôn đặt ra các nội dung nghiên cứu cả theo khía cạnh khoa học lẫn theo khía cạnh công nghệ và triển khai. Tiếp cận tích hợp ontology sử dụng học máy là một tiếp cận được định hướng trong luận án. Qua nghiên các tài liệu trên, các kỹ thuật tích hợp mô hình dữ liệu bao gồm các kỹ thuật chính: tích hợp dữ liệu dựa trên lược đồ dữ liệu, dựa trên thể hiện, dựa trên Ontology, dựa trên việc sử dụng học máy, dựa trên việc sử dụng các độ đo và dựa trên kết hợp một số các kỹ thuật trên với nhau.

1.2. GIỚI THIỆU CHUNG VỀ ONTOLOGY

1.2.1. Khái niệm và phân loại

Theo thời gian, khái niệm của ontology đã được tiến hóa nhằm phù hợp với phạm vi nghiên cứu và triển khai liên quan. Robert Arp và cộng sự [Arp15] giới thiệu một định nghĩa có tính phổ quát về ontology, theo đó "*ontology được định nghĩa là một sản phẩm trình diễn nhân tạo, bao gồm phần đặc thù là một bảng phân loại, trong đó các biểu diễn của nó nhằm chỉ rõ một tổ hợp nào đó của các kiểu, các lớp được định nghĩa và một số quan hệ giữa chúng*".

1.2.2. Thi hành ontology trên hệ thống máy tính

Luận án này tập trung vào việc thi hành ontology trên hệ thống máy tính. M.-A. Sicilia và A. Sicilia [Sicilia14] cung cấp một phác thảo tiến hóa định nghĩa ontology được thi hành trên hệ thống máy tính. Các thành phần chính của ontology thi hành trên hệ thống máy tính gồm: lớp, thực thể, thuộc tính, và các quan hệ.

1.2.3. Nguyên tắc và các bước thiết kế ontology miền

Phần này trình bày về tám nguyên tắc và một quá trình năm bước thiết kế một ontology miền [Arp15].

1.3. GIỚI THIỆU CHUNG VỀ KỸ THUẬT TÍCH HỢP ONTOLOGY

1.3.1. Kỹ thuật tích hợp dữ liệu sử dụng học máy

Học máy là một ngành khoa học, nghiên cứu, xây dựng các kỹ thuật trên nền tảng của trí tuệ nhân tạo giúp cho máy tính có thể dự báo kết quả tương lai thông qua quá trình huấn luyện (học) từ các dữ liệu lịch sử. Một trong các khó khăn khi sử dụng học máy khi triển khai trong thực tế là khi tập dữ liệu huấn luyện (dữ liệu dương, dữ liệu đã được gán nhãn) là rất nhỏ và không có dữ liệu âm. Đã có nhiều nghiên cứu về vấn đề này và đã đem lại những kết quả khả quan [Li07,

Li09, Xiao11, Khan14, Li14, Niu16, Kiryo17]. Đi theo xu hướng này, luận án đã xây dựng một mô hình học máy trong đối sánh ontology dựa trên kho ngữ liệu [VNTrinh1], một mô hình học máy mở rộng ontology từ hai nguồn dữ liệu là một từ điển Anh-Việt và Wikipedia tiếng Việt [VNTrinh4].

1.3.2. Kỹ thuật tích hợp dữ liệu sử dụng các độ đo

Để tích hợp dữ liệu, người ta thường sử dụng các độ đo (measure) để so sánh sự tương đồng giữa các dữ liệu như: Levenshtein, Google, và Cosine [Cohen13]. Một mô hình tích hợp ontology dựa trên các độ đo để đối sánh từ vựng cũng được luận án đề xuất [VNTrinh2].

1.3.3. Kỹ thuật tích hợp dữ liệu sử dụng kết hợp các kỹ thuật trên

Trong bài toán tích hợp dữ liệu, tùy từng bài toán, tùy từng miền dữ liệu, tùy từng bước trong quá trình tích hợp, một số kỹ thuật trên thường được sử dụng kết hợp để tăng cường tính hiệu quả của các thuật toán [Li07, Li09, Bernstein11, Rahm11, Xiao11, Shvaiko13, Khan14, Li14, Niu16, Kiryo17]. Mô hình tích hợp ontology trong [VNTrinh4] được xây dựng dựa trên sự kết hợp kỹ thuật học máy và kỹ thuật dựa trên độ đo.

1.4. CÔNG CỤ TÍCH HỢP DỮ LIỆU VÀ TÍCH HỢP ONTOLOGY

Hầu hết các kỹ thuật đã liệt kê ở trên được cài đặt trong một số lượng lớn các công cụ đối sánh lược đồ dữ liệu và ontology [Rahm11, Euzenat13], như Cupid [Madhavan11], COMA++ [Aumueller05, Do07], ASMOV [Mary09], Falcon-AO [Hu08], RiMON [Li09], AgreementMaker [Cruz09], OII Harmony [Seligman10], [Do02, Bellahsene11], [Euzenat10], [Achichi16]. Phần này nêu nên những điểm mạnh và điểm hạn chế của các công cụ này.

1.5. TÍCH HỢP ONTOLOGY DẦU KHÍ ANH – VIỆT

Nhu cầu tích hợp dữ liệu từ các nguồn dữ liệu khác nhau của Tập đoàn dầu khí Việt Nam (PVN) để xây dựng một hệ thống cung cấp thông tin phục vụ việc ra quyết định một cách chính xác, toàn diện và kịp thời vào hoạt động của Tập đoàn đã trở nên cấp thiết. Do PVN chưa có một ontology chuyên ngành dầu khí, nên việc xây dựng một ontlogy chuyên ngành dầu khí dựa trên các kiến thức đã tổng hợp, nghiên cứu là một việc làm khả thi và hữu ích cho việc tích hợp, và cho việc sử dụng trong công việc chuyên môn, quản lý. Hơn nữa, ứng dụng ontology dầu khí được xây dựng trong các ứng dụng trí tuệ nhóm

(collective intelligence) cũng được đề cập [VNTrinh3]. Từ những lý do trên, một nội dung nghiên cứu - triển khai được định hướng trong luận án là tích hợp dữ liệu để xây dựng ontology dầu khí Anh - Việt.

1.6. KẾT LUẬN CHƯƠNG 1

Chương 1 đã trình bày những nội dung khái quát về tích hợp dữ liệu, tích hợp lược đồ dữ liệu, ontology và tích hợp ontology. Luận án cũng giới thiệu các nguyên tắc thiết kế và các bước triển khai thiết kế một ontology miền. Các kỹ thuật tích hợp mô hình dữ liệu và tích hợp ontology miền đã được trình bày một cách khái quát. Đồng thời, luận án cũng chỉ dẫn các mô hình tích hợp ontology được luận án tập trung nghiên cứu cũng như việc ứng dụng các kết quả nghiên cứu đó vào việc xây dựng ontology dầu khí Anh-Việt tại Viện dầu khí Việt Nam. Các chương tiếp theo sẽ trình bày một cách chi tiết các nghiên cứu của luận án như được chỉ dẫn ở Chương 1.

CHƯƠNG 2. MỘT MÔ HÌNH TÍCH HỢP ONTOLOGY DỰA TRÊN ĐỘ ĐO KHOẢNG CÁCH GOOGLE

2.1. ĐỘ ĐO KHOẢNG CÁCH GOOGLE

2.1.1. Độ phức tạp Kolmogorov

Độ phức tạp Kolmogorov của một xâu *x*, ký hiệu là *K(x)*, được định nghĩa là độ dài tính theo bít của chương trình ngắn nhất sinh ra xâu *x* trên một hệ thống lập trình được tham chiếu. Độ phức tạp Kolmogorov *K(x)* cung cấp giá trị giới hạn dưới của các chương trình sinh ra *x*. Đó là độ dài của chương trình "lý tưởng" sinh ra xâu *x* trong một hệ thống lập trình cụ thể. Trở lại ví dụ trên, K(*x*) là giá trị độ dài nhỏ nhất của xâu kết quả khi nén *x* bằng mọi thuật toán nén có thể.

2.1.2. Khoảng cách thông tin

Cho hai xâu *x* và *y*, *δ* là chương trình ngắn nhất chuyển đổi các xâu sao cho *δ(x) = y* và *δ(y) = x*, độ dài của chương trình δ được gọi là khoảng cách thông tin giữa *x* và *y*. Khoảng cách thông tin giữa x và y, được ký hiệu là *E(x, y)*, được tính theo công thức [Li97]:

$$E(x, y) = K(x, y) + min\{K(x), K(y)\}$$

trong đó *K(x, y)* là độ dài của chương trình nhỏ nhất sinh ra cặp *x, y* và cách để phân biệt chúng.

Khoảng cách thông tin chuẩn hóa *(Normalized Information Distance - NID)* của hai xâu x và y, ký hiệu là *NID (x, y)*, là một hàm khoảng cách thông tin có giá trị thuộc [0, 1] khi xét đến độ dài của các xâu đầu vào. Công thức tính khoảng cách *NID (x, y)* như sau:

$$NID(x,y) = \big(K(x,y) - \min\big(K(x),K(y)\big)\big)/\big(\max\big(K(x),K(y)\big)\big)$$

Gọi *C* là một hàm nén và *C(x)* trả kết quả là xâu được nén của *x*, khi đó khoảng cách nén chuẩn hóa được định nghĩa như sau:

$$NCD_C(x,y) = \big(C(x,y) - \min\big(C(x),C(y)\big)\big)/\max\big(C(x),C(y)\big)$$

2.1.3. Độ đo Google và tính chất

R. Cilibrasi và P. M. B. Vitányi đề xuất các độ đo khoảng cách Google [Cilibrasi4a, Cilibrasi07] thay thế các độ đo khoảng cách nén trong việc xấp xỉ khoảng cách thông tin. Thay vì sử dụng các hàm nén trong các độ đo khoảng cách nén, các độ đo khoảng cách Google sử dụng thông tin được cung cấp từ hệ thống tìm kiếm Google.

Với một xâu *x*, độ phức tạp *C(x)* sẽ trả lại độ dài của kết quả nén xâu x bởi hàm nén *C*. Trong khi đó mã Google của độ dài *G(x)* biểu diễn độ dài từ có mã ngắn nhất được mong đợi của biến cố e_x. Giá trị kỳ vọng này nhận được từ phân phối Google g. Do đó, phân phối Google được sử dụng như bộ nén cho ngữ nghĩa Google. Kết hợp với họ các hàm khoảng cách nén được chuẩn hóa ở trên, khoảng cách Google chuẩn hóa NCD_G (Normalized Compress Distance) được định nghĩa như sau:

$$NCD_G(x,y) = \frac{G(x,y) - \min\big(G(x),G(y)\big)}{\max\big(G(x),G(y)\big)} \qquad (5)$$

Kết hợp công thức (5) với các công thức (1), (2), (3) và (4) ở trên và thực hiện một số biến đổi đơn giản, nhận được:

$$NCD_G(x,y) = \frac{max(log|e_x|, log|e_y|) - log|e_x \cap e_y|}{\log N - min(\log|e_x|, \log|e_y|)} \qquad (6)$$

Đây chính là độ đo khoảng cách Google chuẩn hóa đối với hai xâu *x*, *y*.

Tính chất 1. Khoảng giá trị của NCD_G từ 0 đến $+\infty$.

Tính chất 2. NCD_G là một khoảng cách nhưng không là metric.

2.2. MỘT MÔ HÌNH TÍCH HỢP ONTOLOGY THEO TỪ VỰNG DỰA TRÊN ĐỘ ĐO KHOẢNG CÁCH GOOGLE

Luận án xem xét một phương án đối sánh từ vựng sử dụng độ đo Google và sau đó tích hợp hai ontology miền.

2.2.1. Phát biểu bài toán

Cho hai ontology miền O_1, O_2 về cùng một miền đang được quan tâm. Mỗi ontology O_1, O_2 chứa một tập các khái niệm tương ứng. Mỗi khái niệm này có thể bao gồm tập các thuộc tính; hiển nhiên rằng các thuộc tính của một khái niệm trong cùng một ontology là phân biệt

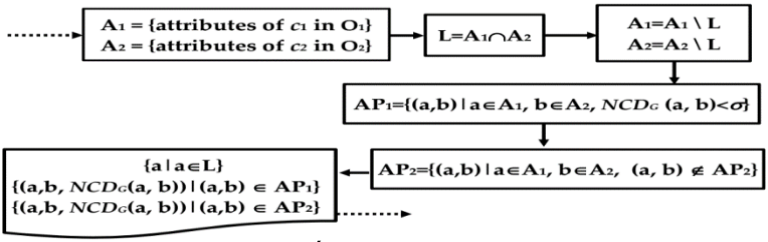
nhau. Lưu ý rằng, cùng một khái niệm ở trong hai ontology O_1, O_2 có thể có số lượng thuộc tính khác nhau.

Bài toán thứ nhất: Cho hai khái niệm khái niệm $c_1 \in O_1$ và khái niệm $c_2 \in O_2$ hãy đối sánh các thuộc tính của khái niệm c_1 và c_2.

Bài toán thứ hai: Cho khái niệm $c_1 \in O_1$ và khái niệm $c_2 \in O_2$, hãy đối sánh hai khái niệm này.

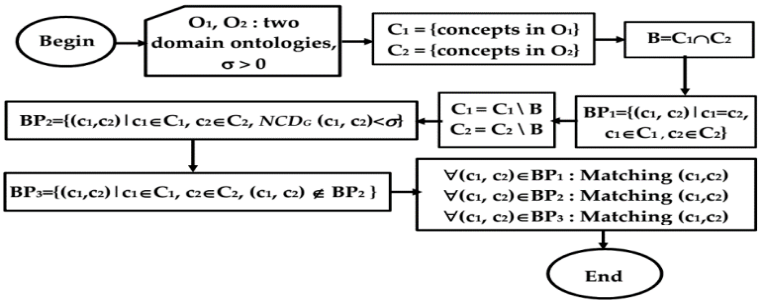
2.2.2. Mô hình đối sánh các thuộc tính của hai khái niệm thuộc hai ontology miền

Hình 2.1 chỉ dẫn mô hình giải quyết bài toán thứ nhất: đối sánh các thuộc tính của hai khái niệm thuộc hai ontology miền.



Hình 2.1. Mô hình đối sánh hai khái niệm thuộc hai ontology miền dựa trên các thuộc tính (Procedure Matching (c1, c2)).

2.2.3. Mô hình đối sánh các khái niệm và tích hợp hai ontology dựa trên độ đo khoảng cách Google



Hình 2.2. Mô hình đối sánh các khái niệm thuộc hai ontology miền

2.2.4. Thực nghiệm

Trong phần này chúng ta xem xét một ứng dụng của độ đo Google đó là dùng để đối sánh các ontology của một bộ truyền áp suất

được sử dụng trong khai thác dầu khí với thông tin phần tiêu đề (khái niệm) của hai ontology Norsock O_1 và ShareCat O_2. Các thuộc tính của ShareCat gồm có: *Document Number, Revision, Plant/Platform, Process Datash. No., Tag number, SerialNo, Range From, SetPoint Low, Range To, SetPoint Height, Range Unit, P&ID, Area, Line/Equipment no., Service description* và các thuộc tính của Norsock gồm có: *Tag number, Scale Range, Service description, Set/Alarm Point, P&ID, Area, Line / equipment no., P. O. Number.* Kết quả thực hiện lược đồ đối sánh được đề xuất bao gồm:

- L = {*Area, Line/equipment no., P&ID, Service description, Tag number*},
- Ma trận khoảng cách Google giữa các khái niệm này được tính như trong Bảng 2.1. Với giá trị σ = 0.2, nhận được tập AP1 = {(*Process Datash. No., Set/Alarm Point*), (*Process Datash. No., P. O. Number*)}.
- Tập thuộc tính L, các cặp thuộc tính trong AP1 và các cặp thuộc tính còn lại (AP2) trong Bảng 2.1 cùng với độ đo khoảng cách Google chuẩn của chúng được hiển thị. Kết quả này cung cấp một gợi ý đối sánh các thuộc tính của cùng một khái niệm trong hai ontology.

Bảng 2.1. Ma trận khoảng cách giữa các thuộc tính trong hai ontology

O_1 \ O_2	*Scale Range*	*Set/Alarm Point*	*P. O. Number*
Document Number	0.5822	0.6998	0.2390
Revision	0.7572	0.8403	0.4187
Plant/Platform	0.7391	0.3959	0.3564
Process Datash. No.	0.4956	**0.1678**	**0.0757**
SerialNo	0.7961	0.5603	0.4692
Range From	0.6055	0.7736	0.4852
SetPoint Low	0.5051	0.3176	0.2859
Range To	0.5679	0.7494	0.4312
SetPoint Height	1.0000	1.0000	1.0000
Range Unit	0.6545	0.5524	0.4973

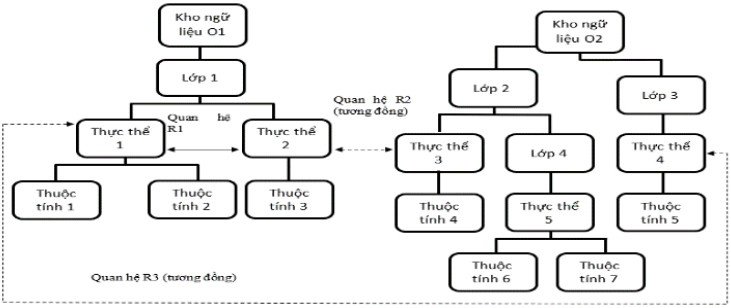
2.3. KẾT LUẬN CHƯƠNG 2

Chương này đã trình bày cơ sở lý thuyết về độ đo Google, bao gồm: độ phức tạp Kolmogorov, khoảng cách thông tin được chuẩn hóa, khoảng cách nén được chuẩn hóa, khoảng cách Google, phân bố

xác suất, ngữ nghĩa và công thức tính của độ đo Google cũng như các tính chất. Một mô hình tích hợp được đề xuất là mô hình đối sánh các khái niệm thuộc hai ontology miền và đối sánh các thuộc tính của hai khái niệm sử dụng độ đo Google. Một trong số các ứng dụng tiêu biểu của độ đo Google để đối sánh các thuộc tính và đối sánh các khái niệm thuộc hai ontology miền dầu khí được giới thiệu. Kết quả nghiên cứu về độ đo Google này đã được trình bày trong [VNTrinh2, VNTrinh5]. Độ đo Goolge đã được ứng dụng để tích hợp dữ liệu trong bài toán mở rộng Ontology Dầu khí Tiếng Việt [VNTrinh4], trong đó, độ đo khoảng cách Google được ứng dụng để tính toán độ tương đồng giữa các khái niệm Tiếng Việt của Từ điển Dầu khí ANH-VIỆT với các khái niệm trong Wikipedia Tiếng Việt.

CHƯƠNG 3. MỘT MÔ HÌNH TÍCH HỢP ONTOLOGY TỪ TÀI NGUYÊN KHO NGỮ LIỆU DỰA TRÊN HỌC MÁY MAXIMUM ENTROPY VÀ BEAM SEARCH

3.1. MÔ HÌNH TÍCH HỢP ONTOLOGY DỰA TRÊN CÁC KHO NGỮ LIỆU SỬ DỤNG PHƯƠNG PHÁP HỌC MÁY MAXIMUM ENTROPY VỚI BEAM SEARCH



Hình 3.1. Mô hình tích hợp ontology dựa trên các kho ngữ liệu sử dụng Phương pháp học máy

Thuật toán học máy được đề xuất trong mô hình này chính là Maximum Entropy và Beam Search. Việc sử dụng phương pháp Maximum Entropy và Beam Search này là hợp lý vì nó có thể ***huấn luyện một số lượng lớn các đặc trưng và hội tụ nhanh*** [Berger96],

[Borthwick98], [McCallum00], [Ratnaparkhi96]. Do độ phức tạp tính toán lớn hơn độ phức tạp tính toán của thuật toán Beam Search O(kT), nên thuật toán Beam search được chọn và sử dụng trong luận án.

Để kiểm chứng về mô hình đề xuất, một ứng dụng của mô hình trên đã được áp dụng trong miền dữ liệu y sinh và đã mang lại kết quả khả quan. Kết quả của việc áp dụng mô hình tích hợp ontology từ tài nguyên các kho ngữ liệu sử dụng học máy Maximum Entropy với Beam Search trong miền dữ liệu y sinh được trình bày dưới đây và đã được công bố trong [VNTrinh1]. Mô hình này, cũng đã được áp dụng để tích hợp thành công ontology dầu khí Anh – Việt bằng cách sử dụng các kho ngữ liệu Wordnet và kho ngữ liệu Wikipedia Tiếng Việt. Kết quả được công bố trên [VNTrinh4, VNTrinh5].

3.2. ÁP DỤNG MÔ HÌNH TÍCH HỢP ONTOLOGY SỬ DỤNG CÁC KHO NGỮ LIỆU TRONG MIỀN DỮ LIỆU Y SINH

3.2.1. Tập ngữ liệu biểu hiện

Mục đích xây dựng một tập dữ liệu cho nhận dạng thực thể biểu hiện với điều kiện là tập dữ liệu thử nghiệm và dữ liệu huấn luyện tương đối nhỏ và được rút ra từ các lĩnh vực gần. Để làm được điều này, ba tập dữ liệu đã được sử dụng: (1) hai tập dữ liệu Phenominer về các bệnh tự miễn dịch và bệnh tim mạch trong công việc [Collier14], (2) một tập dữ liệu trong công việc [Khordad11], tất cả đều được chọn từ các bài tóm tắt Medline trong PubMed đã được trích dẫn bởi các chuyên gia về công nghệ sinh học trong cơ sở dữ liệu về các bệnh di truyền, the Online Mendelian Inheritance of Man (OMIM) [Hamosh05].

3.2.2. Mô hình Maximum Entropy với Beam Search

Tương tự như [Collier13], một phương pháp học máy phù hợp gọi là mô hình Maximum Entropy với Beam Search đã được sử dụng trong nghiên cứu này. Việc sử dụng phương pháp này là hợp lý vì nó có thể ***huấn luyện một số lượng lớn các đặc trưng và hội tụ nhanh***. Sự đánh giá của mô hình này là để đánh giá sự khác biệt nhỏ nhất có thể với thông tin cho trước. Để cài đặt Maximum Entropy với Beam Search, công cụ OpenNLP[3] viết bằng Java với các tham số mặc định đã được sử dụng. Để huấn luyện mô hình nhận dạng thực thể kiểu

[3] http://opennlp.apache.org/

hình, một số đặc trưng và nguồn tài nguyên bên ngoài (các từ điển, ontology) được sử dụng, như ***Human Phenotype Ontology*** **(HPO) [Robinson08] và *Mammalian Phenotype Ontology* [Smith04].**

Quá trình xây dựng

Thứ nhất, tiến hành xây dựng một tập dữ liệu huấn luyện để xác định các thực thể biểu hiện ở người. Bằng cách kết hợp hai mối quan hệ (mối quan hệ giữa các thuật ngữ trong HPO và các tài liệu từ cơ sở dữ liệu OMIM trích ra từ tập tin Phenotype annotation.tab và mối quan hệ giữa mỗi tài liệu của cơ sở dữ liệu OMIM và các tóm tắt Pubmed), đã tập hợp các mối quan hệ giữa các thực thể biểu hiện liên quan đến tóm tắt Pubmed ở con người và các thuật ngữ của HPO. Thu thập tất cả các tóm tắt trong danh sách mối quan hệ ở trên, tùy thuộc vào mỗi bản tóm tắt được tham chiếu đến một danh sách riêng các thuật ngữ HPO từ tập tin mối quan hệ, sử dụng một phương pháp có tên Noun Chunking để gắn nhãn các thực thể biểu hiện trong mỗi tóm tắt. Phương pháp Noun Chunking tìm tất cả các danh từ và cụm từ danh từ trong mỗi tóm tắt Pubmed và so sánh chúng với một danh sách riêng biệt mà tham chiếu đến một số thuật ngữ biểu hiện HPO cụ thể để gán nhãn. Cuối cùng, đã thu được tập dữ liệu HPO NC theo phương pháp này.

Một tập dữ liệu huấn luyện cũng đã được xây dựng để xác định các thực thể biểu hiện ở động vật có vú. Thứ nhất, thu thập mối quan hệ giữa mỗi tóm tắt Pubmed liên quan đến các thuật ngữ trong ontology MP từ hai tệp thống kê: MGI GenoPheno.rpt và MGI PhenoGenoMP.rpt. Nhóm các bài tóm tắt Pubmed trong danh sách mối quan hệ trên, tùy thuộc vào mỗi bản tóm tắt được tham chiếu đến một danh sách riêng các thuật ngữ MP, cũng đã sử dụng Noun Chunking để gắn nhãn thực thể biểu hiện ở động vật có vú cho các bài tóm tắt Pubmed. Một tập dữ liệu huấn luyện MP NC đã được tạo ra như là một kết quả của quá trình trên.

Bước tiếp theo, ghép nối hai tập HPO NC và MP NC để có được tập HPO MP NC với vùng phủ rộng lớn trong miền dữ liệu thực thể biểu hiện.

Bảng 3.1. Thống kê các tập dữ liệu

	HPO_NC	**MP_NC**	**HPO_MP_NC**
Abstracts	18.021	4.035	22.056
Tokens	3.387.015	988.598	4.375.613
Phenotype entities	39.454	6.833	46.287
Unique phenotype entities	3.579	1.169	4.371

Hiệu quả của phương pháp tự động tạo ra tập dữ liệu bằng cách sử dụng phương pháp học máy (ME + BS) với 17 loại đặc trưng trên ba tập dữ liệu huấn luyện chuẩn: Phenominer 2012, Phenominer 2013 và Khordad corpus, đã được đã đánh giá. Bảng 3.4. như là một kết quả của việc đánh giá các kho dữ liệu huấn luyện sinh tự động trên các tập dữ liệu Phenominer 2012 và Phenominer 2013 và Khordad.

Bảng 3.2. Đánh giá các kết quả

Testing data	**Phenominer 2012**			**Phenominer 2013**			**Khordad corpus**		
Training data	P	R	F	P	R	F	P	R	F
HPO_NC	55.37	20.28	29.69	59.82	25.08	**35.34**	89.57	68.21	77.44
MP_NC	40.08	17.44	24.3	42.64	20.78	27.94	83.24	61.09	70.47
HPO_MP_NC	55.69	22.17	**31.71**	58.47	23.97	34	88.12	70.54	**78.36**

Tóm lại, nghiên cứu này đã trình bày một cách có hệ thống về cách xây dựng một tập dữ liệu huấn luyện tự động cho việc nhận dạng thực thể biểu hiện từ các ontology nguồn khác nhau và các phương pháp. Đây là nghiên cứu đầu tiên để đánh giá một tập lớn các đặc trưng cho lớp phức tạp các biểu hiện. Tập dữ liệu được đánh giá bằng cách sử dụng nhận dạng mô hình thực thể biểu hiện gọi là Phương pháp Maximum Entropy với thuật toán Beam Search. Bằng phương pháp này, đã đạt được điểm số F tốt nhất vào khoảng 31,71% đối với Phenominer 2012; 35,34% đối với Phenominer 2013 và 78,36% đối với Khordad.

3.3. KẾT LUẬN CHƯƠNG 3

Chương này của luận án đã trình bày một mô hình tích hợp ontology dựa trên các kho ngữ liệu. Trong mô hình này, các thông tin về khái niệm, thuộc tính của các ontology miền (kho ngữ liệu) đã được tích hợp sử dụng các thuật toán học máy và đối sánh từ vựng. Để kiểm chứng tính khả thi của mô hình đề xuất, mô hình đã được áp dụng thử nghiệm vào trong miền dữ liệu y sinh, để xây dựng một tập dữ liệu huấn luyện tự động cho việc nhận dạng thực thể biểu hiện từ các ontology miền khác nhau. Phương pháp Maximum Entropy với thuật toán Beam Search đã được sử dụng. Một phần kết quả nghiên cứu trong chương này đã được công bố trong [VNTrinh1]. Với kết quả được kiểm chứng là tốt, mô hình này đã được dùng để tích hợp từ điển dầu khí Anh-Việt, Ontology Wordnet, Wikipedia Tiếng Việt để xây

dựng ontology dầu khí Anh - Việt và kết quả nghiên cứu đã được công bố trong [VNTrinh4, VNTrinh5], và được trình bày trong chương 5.

CHƯƠNG 4. MỘT MÔ HÌNH TÍCH HỢP ONTOLOY DỰA TRÊN HỌC MÁY VỚI DỮ LIỆU DƯƠNG VÀ DỮ LIỆU CHƯA GẮN NHÃN

4.1. ĐẶT VẤN ĐỀ

Các thuật toán học máy được ứng dụng hiệu quả trong rất nhiều các lĩnh vực, trong đó có tích hợp dữ liệu, tích hợp ontology. Tuy nhiên, một trong những khó khăn đó là khi các dữ liệu dương dùng để huấn luyện mô hình có ít hoặc rất ít. Việc gán nhãn thủ công tốn rất nhiều thời gian và công sức của các chuyên gia. Đến nay, Việt Nam chưa có Ontology dầu khí mà mới chỉ có từ điển dầu khí Anh - Việt. Trong Wikipedia Tiếng Việt có nhiều khái niệm dầu khí.

Luận án sẽ nghiên cứu, đề xuất một mô hình tích hợp ontology dựa trên các thuật toán học máy với dữ liệu dương và dữ liệu chưa gán nhãn, áp dụng vào việc tích hợp dữ liệu từ điển dầu khí Anh - Việt và Wikipedia Tiếng Việt để Xây dựng ontology dầu khí Anh - Việt với số lượng khái niệm dầu khí Tiếng Việt được mở rộng.

4.2. PHÁT BIỂU BÀI TOÁN

Cho một từ điển dầu khí Tiếng Việt bao gồm một tập các khái niệm dầu khí cùng với các giải thích của chúng. Cho Wikipedia Tiếng Việt trong đó có lĩnh vực dầu khí. Bài toán đặt ra là tích hợp dữ liệu từ hai nguồn dữ liệu trên.

4.3. MÔ HÌNH ĐỀ XUẤT

Hình 4.3 trình bày mô hình đề xuất cho việc tích hợp dữ liệu. Quy trình bao gồm hai giai đoạn như mô tả dưới đây.

4.3.1. Hai giai đoạn tích hợp dữ liệu

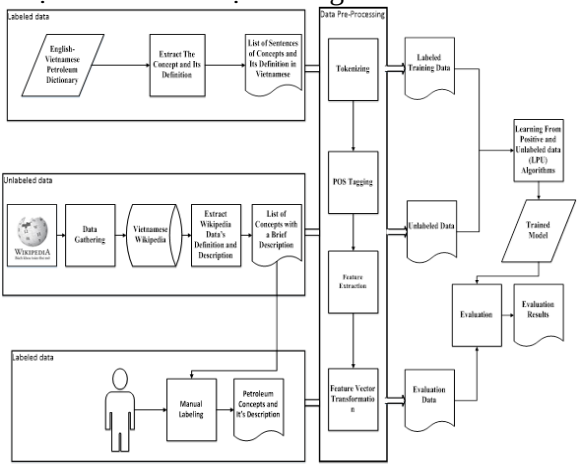
Giai đoạn 1. Lọc các khái niệm trong miền dữ liệu tiềm năng

Tích hợp dữ liệu dựa trên đối sánh từ vựng bằng cách sử dụng đối sánh từ vựng trực tiếp các khái niệm giữa hai tập dữ liệu.

Bước 1. Đối sánh từ vựng trực tiếp từng khái niệm trong số 11.139 khái niệm với từng khái niệm trong 7.155.700 khái niệm trong Wikipedia Tiếng Việt để trích chọn ra những khái niệm chung.

Bước 2. Từ các khái niệm của từ điển và Wikipedia Tiếng Việt, tách thành các từ, cụm từ có nghĩa, xóa bỏ các từ dừng, từ vô nghĩa. Xây dựng các đặc trưng và vectơ đặc trưng.

Bước 3. Đối sánh từ vựng trực tiếp từng khái niệm (đã được đặc trưng hóa) ở trên với các khái niệm trong Wikipedia Tiếng Việt để trích chọn ra các khái niệm chung.



Hình 4.1. Mô hình tích hợp dữ liệu đề xuất cho Ontology dầu khí

Giai đoạn 2. Đối sánh khái niệm

Đối sánh khái niệm dựa trên đối sánh gián tiếp các khái niệm của hai nguồn dữ liệu sử dụng học với dữ liệu dương và dữ liệu chưa gán nhãn kết hợp với độ đo khoảng cách Google và độ đo khoảng cách Cosine để tính sự tương tự giữa mô tả của các khái niệm trong hai tập dữ liệu từ điển và Wikipedia.

$SIM_{Total}(C1, C2) = \alpha * NCD_{Cosine}(C1, C2) + (1 - \alpha) * NCD_G(C1, C2)$

Trong đó, SIM_{Total} là độ tương tự cuối cùng, C1 and C2 là khai niệm cần đối sánh. NCD_{Cosine} là độ đo khoảng cách Cosine. NCD_G là độ đo khoảng các Google chuẩn.

4.3.2. Các thành phần chính

Thành phần xử lý dữ liệu Wikipedia, Thành phần tiền xử lý dữ liệu (Data pre-processing component), Thành phần phân lớp dữ liệu (Data classification component), và Thành phần tạo tập dữ liệu đánh giá (Evaluation dataset construction component). Chiến lược hai bước được sử dụng để giải quyết bài toán này. Một cấu trúc phân tầng khái niệm theo các độ đo được áp dụng. Tại bước thứ nhất, tập dữ liệu âm "tin cậy" ("reliable" negative (RN)) phải được xác định. Tại bước thứ

hai, một bộ phân lớp tốt dựa trên phương pháp lặp sẽ được xây dựng và chọn lựa. Trong luận án này, ba thuật toán sẽ được cài đặt, gồm PERL, ROC-SVM, và DISTANCE. Công cụ LPU [Li07] được sử dụng để chạy các thuật toán PERL và ROC-SVM.

4.4. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

4.4.1. Dữ liệu thực nghiệm

Từ điển dầu khí Anh - Việt, Wikipedia Tiếng Việt, Dữ liệu đánh giá. Kết quả thực nghiệm trên ba độ đo P, R, F trên dữ liệu dương.

4.4.2. Các trường hợp thực nghiệm

Có 3 thực nghiệm được thực hiện trong nghiên cứu này.

4.4.3. Kết quả thực nghiệm

Phần này trình bày kết quả của các thực nghiệm. Kết quả của thí nghiệm 2 được trình bày trong Bảng 4.1 và kết quả của thí nghiệm 3 được trình bày trong Bảng 4.2.

Bảng 4.1. Kết quả các độ đo P, R, F của các thuật toán

Method		*P*	*R*	*F*
PERL		80.24	76.36	78.25
ROC/ISVM	Cosine	82.53	79.21	80.84
	NCD_G	67.08	70.45	68.72
DISTANCE/ISVM	Cosine	84.17	80.49	82.29
	NCD_G	73.25	75.61	74.41

Bảng 4.2. Sự phụ thuộc của độ đo F trong thuật toán ROC/ISVM và DISTANCE vào tỷ lệ α

α / Method	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
ROC/ISVM (Hybrid)	68.72	72.59	75.67	76.88	78.49	80.36	82.35	82.41	80.57	81.29	80.84
DISTANCE (Hybrid)	74.41	79.34	80.46	81.53	82.79	83.41	83.17	81.56	82.67	82.19	82.29

4.4.4. Kết quả xây dựng Ontology dầu khí Tiếng Việt

Khi áp dụng thuật toán phân lớp với dữ liệu dương và dữ liệu chưa gán nhãn kết hợp với độ đo Google và Cosine với tỷ lệ $\alpha = 0.50$ để tích hợp các khái niệm giữa từ điển dầu khí với Wikipedia Tiếng Việt chúng ta thu được 5.084 khái niệm dầu khí, được các chuyên gia

dầu khí của Viện Dầu khí Việt Nam đã kiểm tra sơ bộ và đánh giá cao.

4.4.5. Nhận xét đánh giá

Từ các kết quả thực nghiệm ở trên, chúng ta thấy rằng: (1) Phương pháp dựa trên khoảng cách cho kết quả tốt hơn các phương pháp còn lại; (2) Độ đo khoảng cách Cosine tốt hơn NCD_G do dựa trên đặc trưng mô tả của hai khái niệm; (3) Việc kết hợp hợp Cosine và NCD_G giúp tăng độ chính xác của kết quả với tham số trộn alpha = 0.5 đối với phương pháp Distance và 0.7 đối với phương pháp ROC/ISVM. (4) Tích hợp dữ liệu sử dụng thuật toán phân lớp với dữ liệu dương và dữ liệu chưa gán nhãn kết hợp với độ đo Google và độ đo Cosine (giai đoạn 2 của mô hình đề xuất) thì tốt hơn thuật toán đối sánh từ vựng trực tiếp (giai đoạn 1 của mô hình đề xuất). (5) Một Ontology Dầu khí Tiếng Việt mới hoàn toàn được sinh ra với 16.084 khái niệm, tăng 5.084 khái niệm Tiếng Việt so với từ điển ban đầu.

4.5. PHÁT TRIỂN MÔ HÌNH ĐỀ XUẤT

4.5.1. Giới thiệu

Từ điển Anh-Việt chuyên ngành dầu khí gồm có hơn 11 nghìn khái niệm liên quan đến dầu khí, như đã trình bày ở mục 1.5. Nguồn dữ liệu Wordnet, với 114.000 khái niệm tiếng Anh, trong đó các khái niệm dầu khí, liên kết với nhau thông qua một số mối quan hệ. Có tất cả 20 loại quan hệ giữa các khái niệm trong ontology Wordnet. Bài toán đặt ra là tích hợp hai nguồn dữ liệu trên để xây dựng một ontology dầu khí Anh -Việt, có cấu trúc, và có chứa các mối quan hệ về mặt ngữ nghĩa giữa các khái niệm, hoạt động trên nền tảng web-based, với giao diện đồ họa thân thiện, dễ sử dụng. Hiện nay, có nhiều công cụ được sử dụng để hỗ trợ trong việc xây dựng Ontology. Công cụ Protégé là công cụ được đánh giá là tốt nhất, tuy nhiên công cụ này vẫn còn yếu điểm là không hỗ trợ việc thêm một ontology mới (kế thừa) và hạn chế trong việc hỗ trợ đa người dùng (phân cấp phân quyền, cộng tác) [Khondoker10], [GFC07].

4.5.2. Phương pháp

Việc tích hợp dữ liệu giữa hai nguồn trên để xây dựng ontology dầu khí được mô tả như sau.

Bước 1. Sử dụng phương pháp đối sánh từ vựng so sánh một khái niệm Tiếng Anh trong từ điển với một khái niệm Tiếng Anh trong Wordnet Tiếng Anh, để lấy tất cả những khái niệm Tiếng Anh vừa có quan hệ với nó trong Wordnet Tiếng Anh vừa có trong từ điển dầu khí thì lấy ra và cho vào ontology mới, cùng với các mối quan hệ tương ứng của

các khái niệm Tiếng Anh này. *Bước 2*. Từ nguồn dữ liệu từ vựng, các chuyên gia sẽ định nghĩa ra các (lớp) nhóm từ tương ứng với các nhóm lĩnh vực trong ngành công nghiệp Dầu khí. Sau đó, chuyên gia sẽ nhập liệu một số từ mẫu vào các nhóm tương ứng để tạo lập bộ dữ liệu huấn luyện. *Bước 3*. Xây dựng công cụ phần mềm hỗ trợ thực hiện việc rút trích tự động ra các đặc trưng tương ứng với từng nhóm mà chuyên gia đã định nghĩa. *Bước 4*. Từ tập các từ đặc trưng do công cụ đề xuất, các chuyên gia có thể kiểm tra, chọn lọc lại các đặc trưng chính xác và loại bỏ các đặc trưng chưa đúng. *Bước 5*. Để nâng cao tốc độ xử lý và độ chính xác khi phân loại, chúng ta sẽ tiến hành loại bỏ các từ dừng, từ ngắt, từ vô nghĩa. Ở bước này, để có thể loại bỏ các từ vô nghĩa thì ta cần phải tách được các từ trong 1 câu tiếng Việt. Để giải quyết vấn đề này, chúng ta sử dụng công cụ JVNTextPro[4] để thực hiện tách từ tiếng Việt. *Bước 6*. Xây dựng công cụ phân lớp các từ vào các nhóm/lớp tương ứng sử dụng thuật toán học máy. *Bước 7*. Sau đó các chuyên gia sẽ kiểm tra lại kết quả phân lớp trước khi cập nhật vào CSDL để làm giàu cho ontology.

4.5.3. Kết quả

Ontology dầu khí Anh – Việt được xây dựng và công cụ hỗ trợ tích hợp cũng được xây dựng.

4.5.4. Nhận xét

Đối sánh từ vựng, tri thức chuyên gia, khái niệm đồng nghĩa Tiếng Việt, và thuật toán học máy đã được sử dụng để xây dựng ontology dầu khí Anh – Việt với 11.139 khái niệm và các mô tả của nó cúng với 6.382 quan hệ kế thừa từ ontology Wordnet. Ontology dầu khí Anh – Việt này hữu ích cho các cán bộ nhân viên ngành dầu khí trong việc nghiên cứu, tra cứu, biên dịch, đào tạo, tích hợp dữ liệu, và mở rộng trong hiện tại và tương lai. Nó cũng có thể được dùng để tích hợp với Wikipedia Tiếng Việt để mở rộng thêm các khái niệm Tiếng Việt bằng cách sử dụng mô hình học với dữ liệu dương và dữ liệu chưa gán nhãn.

4.6. KẾT LUẬN CHƯƠNG 4

Chương này của luận án đã trình bày một mô hình tích hợp ontology dầu khí sử dụng thuật toán học máy với dữ liệu dương và dữ liệu chưa gán nhãn kết hợp với độ đo Google và độ đo Cosine để nâng cao hiệu quả của việc tích hợp. Đồng thời, luận án cũng đưa ra hai ví

[4] http://jvntextpro.sourceforge.net/

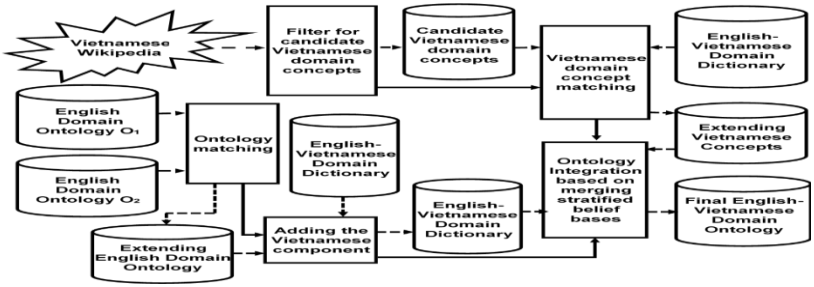
dụ cụ thể để áp dụng mô hình đề xuất trong miền dữ liệu dầu khí, sử dụng từ điển dầu khí Anh – Việt, ontology Wordnet, và Wikipedia Tiếng Việt. Kết quả nghiên cứu về học máy với dữ liệu dương và dữ liệu chưa gán nhãn kết hợp với độ đo Google và độ đo Cosine này đã được trình bày trong [VNTrinh4]. Thuật toán học máy với dữ liệu dương và dữ liệu chưa gán nhãn kết hợp với độ đo Google và độ đo Cosine đã được ứng dụng để tích hợp dữ liệu trong bài toán xây dựng ontology miền dầu khí Anh-Việt được trình bày trong [VNTrinh5] và trong chương 5 của luận án.

CHƯƠNG 5. XÂY DỰNG ONTOLOGY DẦU KHÍ ANH - VIỆT TẠI VIỆN DẦU KHÍ VIỆT NAM

5.1. ĐẶT VẤN ĐỀ

Trên cơ sở các kết quả nghiên cứu được trình bày ở các chương trước trong luận án và nhu cầu thực tiễn của ngành dầu khí Việt Nam, luận án đã tiến hành xây dựng một ontology miền dầu khí, sử dụng kết hợp các thuật toán đã nghiên cứu, phục vụ cho công việc tra cứu, tìm kiếm, nghiên cứu, đạo tạo, dịch thuật hàng ngày của các bộ, nhân viên ngành dầu khí.

5.2. TIẾP CẬN XÂY DỰNG ONTOLOGY QUA NĂM BƯỚC



Hình 5.1. Mô hình xây dựng ontology dầu khí Anh - Việt tại Viện Dầu khí Việt Nam

Hình 5.1 mô tả khung của mô hình tích hợp ontology miền dầu khí Anh – Việt. Khung này bao gồm 5 phần: chính Lọc các khái niệm miền dầu khí Tiếng Việt tiềm năng, Tích hợp ontology, Bổ sung thành phần Tiếng Việt vào ontology miền dầu khí Tiếng Anh, Đối sánh khái niệm miền dầu khí Tiếng Việt, Tích hợp ontology dựa trên tích hợp các cơ sở niềm tin phân tầng. Tiếp cận này cũng đảm bảo tám nguyên tắc thiết kế ontology và năm bước thiết kế ontology. Tiếp cận

xây dựng ontology dầu khí Anh – Việt được thể hiện thông qua một quá trình gồm năm bước sau đây như được thể hiện ở Hình 5.1. ***Bước 1.*** **Lọc các khái niệm miền dầu khí Tiếng Việt tiềm năng.** ***Bước 2.*** **Tích hợp Ontology.** ***Bước 3.*** **Đối sánh khái niệm miền dầu khí Tiếng Việt.** ***Bước 4.*** **Bổ sung thành phần Tiếng Việt vào ontology miền dầu khí Tiếng Anh.** ***Bước 5.*** **Tích hợp ontology dựa trên tích hợp các cơ sở niềm tin phân tầng.** Trong Bước 5, tri thức của các chuyên gia dầu khí đã được sử dụng để kiểm tra, chỉnh sửa các lỗi, chính xác hóa các kết quả của việc tích hợp và việc phân lớp dữ liệu, để nâng cao chất lượng của các kết quả tích hợp. Khi đối sánh hai khái niệm c1 thuộc O1 và c2 thuộc O2, độ đo khoảng cách Google (cơ sở tri thức) trả về một con số (niềm tin) về sự tương đồng giữa c1 và c2, trong khi đó, độ đo khoảng cách Cosine (cơ sở tri thức) cũng trả về một con số khác (niềm tin) về sự tương đồng giữa c1 và c2. Hai kết quả này có thể khác nhau, thậm chí là trái ngược, mâu thuẫn nhau. Ngoài ra, khi sử dụng tri thức chuyên gia dầu khí (cơ sở tri thức) để kiểm tra, rà soát sự tương đồng (niềm tin) giữa hai khái niệm dầu khí c1 và c2, hoặc là khi phân lớp (niềm tin) các khái niệm dầu khí vào các nhóm (lớp) dữ liệu, có thể xuất hiện những mâu thuẫn giữa các chuyên gia. Thuật toán tích hợp ontology dựa trên tích hợp các cơ sở niềm tin phân tầng [VNTrinh3] sẽ giúp giải quyết các bài toán dạng này. Phương pháp tích hợp niềm tin trong tích hợp ontology sử dụng các kỹ thuật tranh luận. Ý tưởng chính là tổ chức mỗi quy trình tích hợp niềm tin như là một trò chơi mà những tác nhân tham gia sử dụng các kỹ thuật tranh luận để tranh luận, dựa trên cơ sở niềm tin của chính họ, để đạt được một sự đồng thuận (một cơ sở niềm tin chung) từ một tình huống mâu thuẫn.

5.3. TRIỂN KHAI

5.3.1. Thu thập và tiền xử lý dữ liệu

Dữ liệu được thu thập từ ba nguồn chính: từ điển dầu khí Anh – Việt, Wordnet[5] Tiếng Anh, và dữ liệu từ Wikipedia[6] Tiếng Việt. Các dữ liệu này được thu thập, chọn lọc, tách câu, tách từ, token hóa, loại bỏ từ dừng, từ nối, từ vô nghĩa. Ngoài ra, danh sách các từ đồng nghĩa Tiếng Việt[7] và danh sách các từ vô nghĩa Tiếng Việt[8] cũng được sử

[5] https://wordnet.princeton.edu
[6] https://wordnet.princeton.edu
[7] http://viet.wordnet.vn
[8] https://github.com/stopwords/vietnamese-stopwords

dụng. Công cụ JVNTextPro[9] , DKPro[10] Java Wikipedia Library, LPU[11] , Thư viện javascript "GoJS"[12] , Microsoft .NET MVC 4.0 (Model-View-Controller), SQL Server 2014 được sử dụng.

5.3.2. Thi hành ontology dầu khí Anh - Việt trên hệ thống máy tính

Áp dụng khung mô hình tích hợp ontology miền dầu khí Anh – Việt, bao gồm 5 bước ở trên.

5.3.3. CÀI ĐẶT

Ontology dầu khí Anh – Việt đã được cài đặt tại máy chủ của Viện Dầu khí Việt Nam.

5.4. KẾT QUẢ

Ontology dầu khí Anh - Việt đã được xây dựng đáp ứng hoàn toàn tất cả các yêu cầu đặt ra, với 11.139 khái niệm dầu khí Tiếng Anh và 16.223 khái niệm dầu khí Tiếng Việt, và các mô tả của chúng trong Tiếng Anh và Tiếng Việt, cùng với 6.823 các mối quan hệ giữa khái niệm thỏa mãn hoàn toàn các yêu cầu đặt ra ban đầu. Biểu diễn đồ họa của các mối quan hệ giữa một khái niệm dầu khí với các khái niệm dầu khí còn lại, và giữa hai khái niệm dầu khí bất kỳ trong ontology dầu khí được thực hiện. Các chức năng quản trị khái niệm và các thông tin liên quan được cài đặt với các giao diện đồ họa. Các công cụ đồ họa hỗ trợ tích hợp dữ liệu cũng được triển khai. Chức năng phân cấp, phân quyền đến từng người dùng và các biện pháp bảo đảm an ninh, an toàn, bảo mật thông tin cũng được thực hiện. Có thể sao lưu, dự phòng và khôi phục dễ dàng. Phần mềm được thiết kế theo hướng mở, tường minh ngay từ trong thiết kế, sử dụng các hệ quản trị cơ sở dữ liệu chuyên nghiệp, thương mại của Microsoft, dễ dàng nâng cấp, mở rộng trong tương lai.

5.5. KẾT LUẬN CHƯƠNG 5

Chương này của luận án đã trình bày một mô hình xây dựng ontology miền dầu khí sử dụng các kết quả nghiên cứu từ các chương khác của luận án như độ đo Google, thuật toán học với dữ liệu dương và dữ liệu chưa gán nhãn, sử dụng các kho ngữ liệu, cơ sở niềm tin phân tầng [VNTrinh3], các nguyên tắc và các bước xây dựng ontology, và nhu cầu thực tiễn của ngành dầu khí Việt Nam. Mô hình này đã

[9] http://jvntextpro.sourceforge.net/
[10] https://dkpro.github.io/dkpro-jwpl/
[11] http://www.cs.uic.edu/~liub/LPU/LPU-download.html
[12] https://gojs.net

được áp dụng để xây dựng một xây dựng một ontology miền dầu khí Anh - Việt cụ thể. Ontology miền dầu khí Anh - Việt này phục vụ hiệu quả cho công việc tra cứu, tìm kiếm, nghiên cứu, đạo tạo, dịch thuật hàng ngày của các bộ, nhân viên ngành dầu khí đáp ứng các yêu cầu về chức năng cũng như yêu cầu thiết kế chuẩn của một ontology.

KẾT LUẬN

I. Những kết quả chính của luận án

Luận án tham gia vào dòng nghiên cứu về tích hợp dữ liệu trên thế giới và đạt được một số đóng góp sau đây về tích hợp dữ liệu.

Thứ nhất, luận án đề xuất bốn mô hình tích hợp dữ liệu. Một là mô hình tích hợp dữ liệu dựa trên độ đo Google [VNTrinh2, VNTrinh4]. Hai là mô hình tích hợp dữ liệu dựa trên các kho ngữ liệu sử dụng học máy Maximum Entropy và Beam Search [VNTrinh1, VNTrinh4, VNTrinh5]. Ba là mô hình tích hợp dữ liệu dựa trên học máy với dữ liệu dương và dữ liệu không gán nhãn [VNTrinh4, VNTrinh5]. Bốn là mô hình tích hợp dữ liệu kết hợp các kỹ thuật trên để xây dựng ontology dầu khí Anh - Việt [VNTrinh1, VNTrinh2, VNTrinh4, VNTrinh5]. Thứ hai, luận án khảo sát ba giải pháp: một là các giải pháp tích hợp dữ liệu dựa trên độ đo, hai là các giải pháp tích hợp dữ liệu dựa trên học máy, ba là giải pháp tích hợp dữ liệu dựa trên các kho ngữ liệu. Thứ ba, trên cơ sở phát triển thuật toán học máy với dữ liệu dương và dữ liệu chưa gán nhãn (Positive and Unlabeled Learning), luận án đã đề xuất một thuật toán kết hợp độ đo Google và độ đo khoảng cách Cosine với thuật toán học máy với dữ liệu dương và dữ liệu chưa gán nhãn để tích hợp dữ liệu, nâng cao hiệu quả của thuật toán. Thứ tư, luận án đóng góp trực tiếp vào hệ thống tích hợp dữ liệu tại Viện Dầu khí Việt Nam. Một ontology Dầu khí ANH - VIỆT được xây dựng dựa trên việc tích hợp từ điển Anh -Việt với Wordnet Tiếng Anh và Wikipedia Tiếng Việt được sử dụng cho việc tra cứu, nghiên cứu, đào tạo trong hiện tại và là cơ sở cho việc mở rộng, tích hợp với các hệ thống dữ liệu khác (ví dụ hệ thống chia sẻ tri thức đang có tại Viện Dầu khí Việt Nam...) và các ontology dầu khí khác trên thế giới, trong tương lai. Các thuật toán tích hợp dữ liệu sử dụng học máy cũng có thể tiếp tục được nghiên cứu để áp dụng cho các bài toán khác trong lĩnh vực thăm dò, khai thác dầu khí (ví dụ: ứng dụng các thuật toán học máy trong tích hợp dữ liệu để nâng cao hệ số thu hồi dầu...). Đồng thời, nhằm minh chứng cho tiềm năng ứng dụng thực tiễn của các mô hình đề xuất, luận án thực thi các thực nghiệm để kiểm chứng tính hữu dụng

của các thuật toán và mô hình được luận án đề xuất. Kết quả thực nghiệm cho thấy tiềm năng ứng dụng cao các kết quả nghiên cứu từ luận án. Luận án cũng có đóng góp trong việc cung cấp một nghiên cứu tổng quan về tích hợp dữ liệu.

II. Hạn chế của luận án

Trong quá trình triển khai các mô hình, luận án vẫn còn tồn tại một số hạn chế như sau: Một là, miền ứng dụng mới áp dụng để xây dựng ontology dầu khí Anh - Việt. Các dữ liệu (khái niệm) chủ yếu ở khâu đầu của chuỗi hoạt động dầu khí, chưa mở rộng ra các khâu khác (khâu giữa, khâu sau). Các dữ liệu rất có giá trị khác liên quan đến hoạt động thăm dò khai thác khác chưa được tích hợp để hỗ trợ ra quyết định (ví dụ: dữ liệu khai thác dầu khí hàng ngày tại các mỏ dầu khí). Hai là, một trong những sản phẩm của luận án là ontology dầu khí Anh - Việt, tuy nhiên, cần phải có thêm thời gian để các chuyên gia dầu khí rà soát, chỉnh sửa, cập nhật để nâng cao chất lượng và độ tin cậy của phần mềm này.

III. Định hướng nghiên cứu tiếp theo

Trong thời gian tiếp theo, nghiên cứu sinh sẽ tiếp tục nghiên cứu các hướng giải quyết cho các hạn chế còn tồn tại của luận án và tiếp tục triển khai các đề xuất để hoàn thiện hơn các giải pháp cho tích hợp dữ liệu. Một là, các kỹ thuật học máy ngày càng được quan tâm cả trong cộng đồng nghiên cứu và ứng dụng, nên sẽ có nhiều các thuật toán mới về học máy và ứng dụng trong tích hợp dữ liệu. Do đó, việc nghiên cứu, áp dụng các thuật toán học máy mới hơn trong tích hợp dữ liệu cũng là một hướng trong tương lai. Hai là, nghiên cứu để phát triển hệ thống hiện có áp dụng kết quả nghiên cứu về tích hợp tri thức. Ba là, nghiên cứu, tìm kiếm, chọn lựa các ontology dầu khí có chất lượng cao trên thế giới để tích hợp với ontology hiện có để mở rộng, tăng thêm số lượng các khái niệm (từ vựng) dầu khí, đặc biệt là các khái niệm thuộc khâu giữa và khâu sau trong chuỗi hoạt động dầu khí. Bốn là, tăng cường sử dụng các tri thức của các chuyên gia dầu khí để kiểm tra, rà soát, chỉnh sửa, bổ sung, để tăng cường tính đúng đắn của các khái niệm, mô tả, quan hệ. Năm là, tích hợp với các hệ thống dữ liệu có sẵn tại Viện dầu khí Việt Nam và Tập đoàn Dầu khí Việt Nam để phát huy hiệu quả của ontology dầu khí này và các hệ thống hiện có (ví dụ: hệ thống quản lý và chia sẻ tri thức tại Viện Dầu khí Việt Nam). Sáu là, tiếp tục nghiên cứu và áp dụng các thuật toán học máy

để tích hợp các dữ liệu trong thăm dò, khai thác, chế biến, lọc hóa dầu, an toàn, môi trường, kinh tế và quản lý dầu khí để hỗ trợ ra quyết định cho lãnh đạo và chuyên gia các cấp, nâng cao hiệu quả sản xuất kinh doanh (ví dụ: ứng dụng học máy để tích hợp dữ liệu khai thác nhằm nâng cao hệ số thu hồi dầu).

DANH MỤC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN TỚI LUẬN ÁN[13]

1. [VNTrinh1] Ngoc-Trinh Vu, Van-Hien Tran, Thi-Huyen-Trang Doan, Hoang-Quynh Le, and Mai-Vu Tran (2015). *A Method for Building a Labeled Named Entity Recognition Corpus Using Ontologies*. Proceedings of 3rd International Conference on Computer Science, Applied Mathematics and Applications - ICCSAMA 2015, pp 141-149. (***Scopus)*.**
2. [VNTrinh2] Trinh Vu Ngoc, Ha Quang Thuy, Tran Trong Hieu. *Độ đo GOOGLE trong tích hợp dữ liệu*. Hội nghị quốc gia lần thứ VIII "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin" (FAIR 2015), trang 224-231. (***Scopus, DBLP)*.**
3. [VNTrinh3] Trong Hieu Tran, Thi Hong Khanh Nguyen, Quang Thuy Ha, and Ngoc-Trinh Vu. *Argumentation framework for merging stratified belief bases*. Asian Conference on Intelligent Information and Database Systems (ACIIDS 2016), pp. 43-53.
4. [VNTrinh4] Ngoc-Trinh Vu, Quoc-Dat Nguyen, Tien-Dat Nguyen, Manh-Cuong Nguyen, Van-Vuong Vu, and Quang-Thuy Ha. *A Positive-Unlabeled Learning Model for Extending a Vietnamese Petroleum Dictionary based on using Vietnamese Wikipedia Data*. ACIIDS (1) 2018: 190-199. (***Scopus, DBLP)*.**
5. [VNTrinh5] Ngoc-Trinh Vu, Hung-Son Nguyen, Quang-Thuy Ha. *An English-Vietnamese Domain Ontology Integration Model and an Application in Oil and Gas Domain*. MAPR 2018 (*submitted*)

[13] **Scopus**: https://www.scopus.com/authid/detail.uri?authorId=56878562200; **DBLP**: http://dblp.uni-trier.de/pers/hd/v/Vu:Ngoc_Trinh