MỞ ĐẦU

Tính cấp thiết của luận án

Ontology (được một số nhà nghiên cứu người Việt gọi là "bản thể học" hoặc "bản thể luận"[1]) là một thành phần tri thức nền tảng và mọi tri thức khác cần được dựa trên và tham chiếu đến nó. Một khu vực ứng dụng ontology vô cùng rộng lớn là trong các công cụ tìm kiếm (search engine) và chia sẻ tri thức (knowledge sharing), ở đó, ontology hỗ trợ đắc lực hoạt động tìm kiếm có cấu trúc, so sánh được và tùy chỉnh cao.

Hiện chưa có ontology tiếng Việt cho miền tài nguyên và môi trường (trong đó bao gồm miền khí hậu Việt Nam), song một vài ontology tiếng Việt cho các miền ứng dụng khác đã được xây dựng, điển hình là ontology VN-KIM, thành phần ontology tiếng Việt trong hệ thống BioCasster và ontology miền dầu khíViệt Nam.

Hiện trên thế giới có một số ontology có liên quan đến miền tài nguyên và môi trường, ví dụ như: SWEET ontology[2](NASA's Semantic Web for Earth and Environment Terminology) là hệ thống các khái niệm về môi trường và trái đất; EnvO ontology[3] (The Environment ontology) là ontology cho miền môi trường và Weather ontology cho miền thời tiết.

Các khái niệm liên quan đến xây dựng ontology thủ công, bán tự động và tự động, bao gồm:

Kỹ thuật ontology (ontology engineering) là việc xây dựng ontology sử dụng các kỹ thuật web ngữ nghĩa thông qua đó lấp đầy cơ sở tri thức (A-Box) với các thể hiện của ontology đó.

Học ontology (ontology learning)là cách *tiếp cận bán tự động xây dựng ontology*bằng việc phát hiện và bổ sung các khái niệm và các quan hệ từ kho văn bản dựa trên việc sử dụng các kỹ thuật khai phá văn bản (text mining) hoặc/và học máy (machine learning). Học ontology là một xu hướng có tính hiện

[1] Do thuật ngữ "bản thể học" hoặc "bản thể luận" là các thuật ngữ có nguồn gốc vay mượn từ ngôn ngữ khác mà không gợi nghĩa nhiều hơn thuật ngữ "ontology" cho nên luận án này sử dụng nguyên gốc "ontology".

[2]https://sweet.jpl.nasa.gov/

[3]http://www.environmentontology.org/

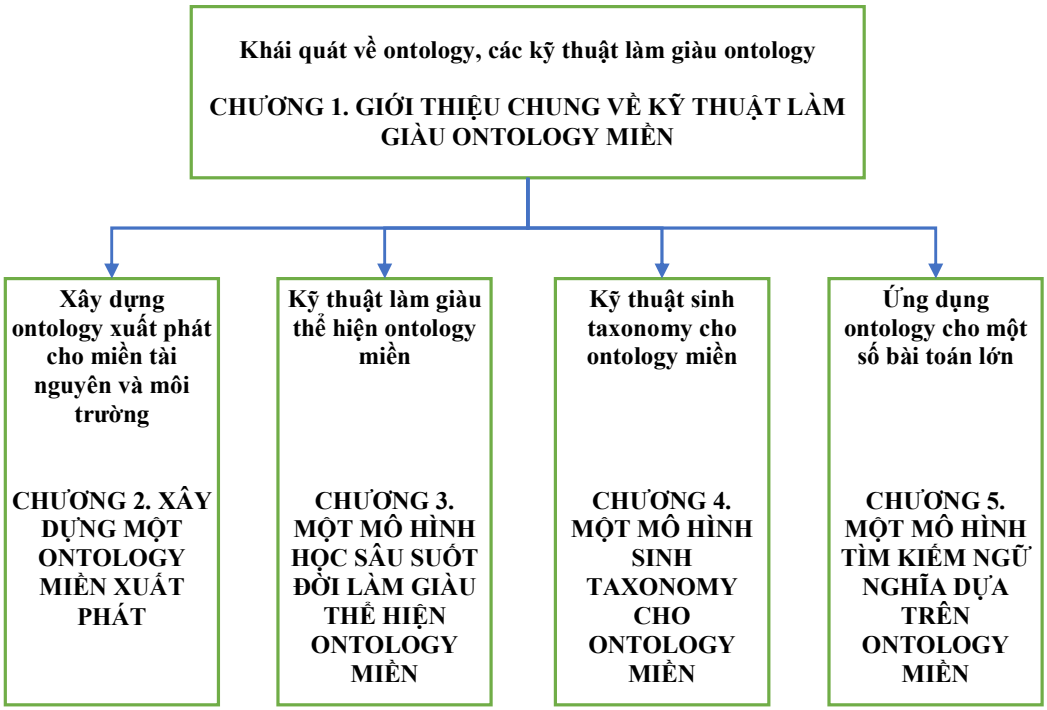
đại và đầy thách thức trong nghiên cứu xây dựng ontology.

Làm giàu thể hiện ontology (ontology population) là cách tiếp cận xây dựng ontology bằng việc phát hiện các thể hiện của các lớp và các thể hiện của các quan hệ và lưu trữ vào trong cơ sở tri thức (A-Box) của ontology [Buitelaar et al., 2005].

Nâng cấp, làm giàu ontology (ontology enrichment) bao gồm các công việc học ontology và làm giàu thể hiện ontology từ một ontology khởi tạo ban đầu.

Với việc hiện nay chưa có ontology cho miền tài nguyên và môi trường và ontology hiện nay được sử dụng rất hiệu quả trong các bài toán tìm kiếm thông tin, xử lý ngôn ngữ, biểu diễn tri thức, ...cũng như tính chất thách thức cao của chủ đề nghiên cứu xây dựng ontology, làm giàu ontology cần các phương pháp bán tụ động dựa trên các kỹ thuật xử lý ngôn ngữ, các kỹ thuật thống kê và các kỹ thuật logic đã tạo động lực nghiên cứu đối với luận án "***Kỹ thuật nâng cấp ontology khí hậu việt nam dựa trên nguồn tài nguyên Web***".

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.



Hình 0.1. Phân bố các chủ đề trong các chương của luận án

CHƯƠNG 1. GIỚI THIỆU CHUNG VỀ KỸ THUẬT LÀM GIÀU ONTOLOGY MIỀN

1.1. GIỚI THIỆU CHUNG VỀ ONTOLOGY

Trong một nỗ lực đưa ra một định nghĩa phổ quát về ontology, R. Arp và cộng sự [1] cho rằng *ontology là một sản phẩm trình diễn do con người tạo ra, với thành phần đặc thù là một bảng phân loại biểu diễn tường minh một tổ hợp nào đó của các kiểu, các lớp được định nghĩa và một số quan hệ giữa chúng*.

Ontology triết học được xây dựng nhằm mục đích cung cấp một phân loại rõ ràng và toàn diện về tất cả các thực thể trong mọi lĩnh vực của cuộc sống.

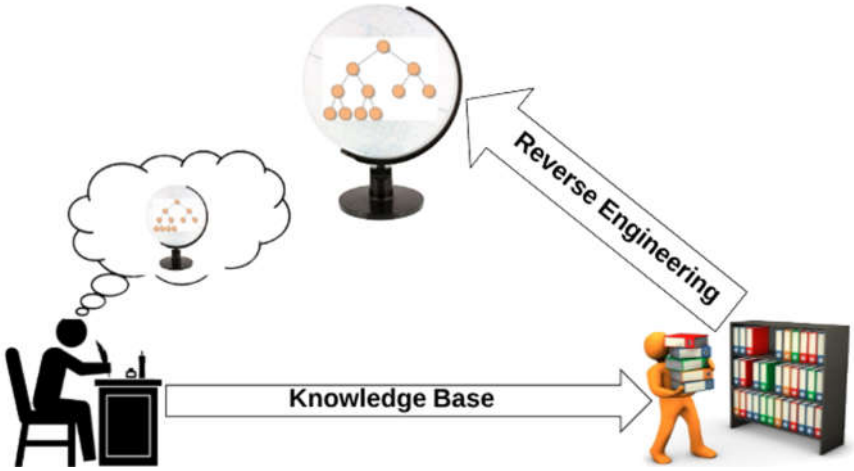
Ontology miền (domain ontology), còn được gọi là ontology cụ thể (material ontology), trình diễn các thực thể và các quan hệ giữa chúng trong một miền thực tiễn cụ thể chẳng hạn như y tế, địa lý, sinh học, luật học nhằm mục đích hỗ trợ trực tiếp các nghiên cứu về lĩnh vực cụ thể được đề cập.

Ontology mức cao (top-level ontology), còn được gọi là ontology hình thức (formal ontology), trình diễn một ontology miền có tính tiêu chuẩn với một kiến trúc phổ quát dùng chung trong cộng đồng, giúp kết nối các ontology khác nhau trong cùng một miền hoặc trong một số miền liên quan nhau.

Ontology ứng dụng (application ontology) được tạo ra nhằm mục đích thực hiện một số bài toán hoặc ứng dụng cụ thể.

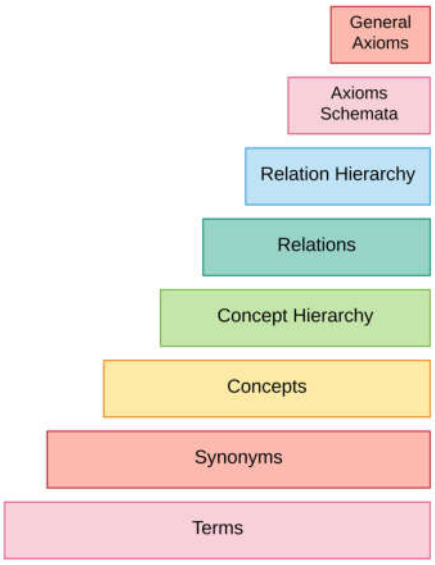
Học ontology

Ontology là cấu trúc chính thức để biểu diễn các khái niệm và các quan hệ của một khái niệm chia sẻ. Chính xác hơn, nó có thể được định nghĩa như các khái niệm, quan hệ, thuộc tính và phân cấp hiện diện trong miền. Tuy nhiên, việc xây dựng các ontology lớn một cách thủ công là một nhiệm vụ khó khăn và việc xây dựng ontology cho tất cả các miền là không khả thi [2]. Do đó, thay vì xây dựng các ontology một cách thủ công, xu hướng nghiên cứu hiện đang chuyển sang học ontology bán tự động hoặc tự động.



Hình 1.1. Học ontology từ văn bản: công việc kỹ thuật đảo ngược[3]

Học ontology là một quá trình ngược lại khi mô hình miền được xây dựng lại từ văn bản đầu vào bằng cách khai thác cấu trúc chính thức được lưu trong tâm trí tác giả. Toàn bộ quá trình xây dựng lại mô hình miền được minh họa trong hình 1.1. Hình 1.2 tóm tắt các bước khác nhau cần thiết để thực hiện xây dựng ontology từ văn bản phi cấu trúc.



Hình 1.2. Các tầng học ontology[3]

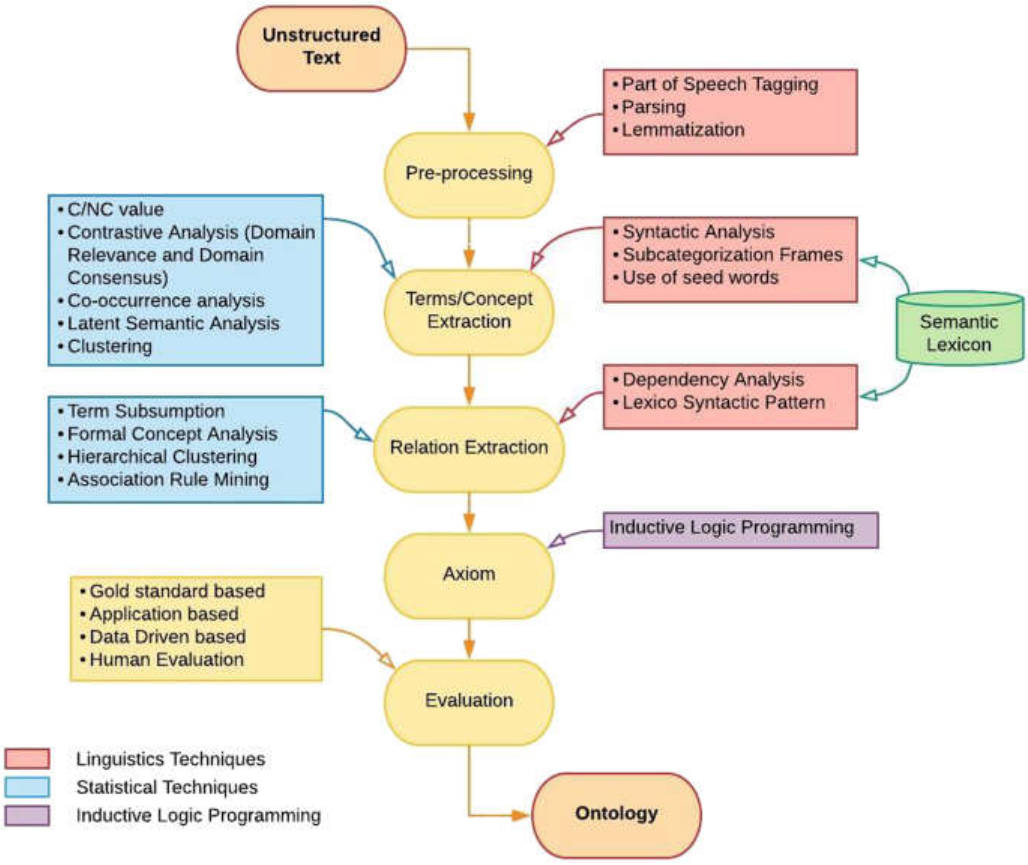
1.2. CÁC KỸ THUẬT LÀM GIÀU ONTOLOGY MIỀN

Các kỹ thuật làm giàu ontology miền hiện được chia làm 3 nhóm chính (hình 1.3):

- Nhóm dựa trên thống kê: bao gồm các kỹ thuật dựa trên học máy, khai phá dữ liệu và tìm kiếm thông tin.

- Nhóm dựa trên xử lý ngôn ngữ: bao gồm các kỹ thuật xử lý ngôn ngữ tự nhiên.

- Nhóm dựa trên logic (Inductive Logic Programming - ILP): bao gồm các kỹ thuật logic mô tả, …



Hình 1.3. Các phương pháp học ontology [3]

Kỹ thuật làm giàu ontology miền dựa trên xử lý ngôn ngữ

Các kỹ thuật dựa trên xử lý ngôn ngữ là các kỹ thuật dựa vào đặc tính của ngôn ngữ và đóng vai trò quan trọng trong mọi pha của quá trình học ontology. Các kỹ thuật dựa trên xử lý ngôn ngữ phần lớn được sử dụng trong quá trình tiền xử lý dữ liệu cũng như trong một vài công việc học ontology khác như trích xuất các thuật ngữ, khái niệm và quan hệ. Như vậy, các kỹ thuật dựa trên xử lý ngôn ngữ có thể chia thành các nhóm:

- Nhóm kỹ thuật phục vụ tiền xử lý dữ liệu, bao gồm: gán nhãn từ loại (part of speech tagging), phân tích cú pháp (parsing) và xác định biến thể từ loại (lemmatization).

- Nhóm kỹ thuật phục vụ trích xuất các thuật ngữ, khái niệm và quan hệ, bao gồm: phân tích phụ thuộc (dependency analysis), phân tích cú pháp từ vựng (lexico-syntactic analysis), phân loại thuật ngữ, phân tích khái niệm hình thức (FCA), khai phá luật kết hợp và phân cụm phân cấp (ARM).

Kỹ thuật làm giàu ontology miền dựa trên thống kê

Các kỹ thuật dựa trên thống kê chỉ dựa trên thống kê của kho văn bản mà không quan tâm đến ngữ nghĩa của chúng. Phần lớn các kỹ thuật thống kê sử dụng nhiều đến các phương pháp xác suất và thường được sử dụng trong các cấp độ đầu tiên của quá trình học ontology sau khi đã tiền xử lý về mặt ngôn ngữ. Các kỹ thuật ngày phần lớn sử dụng cho việc trích xuất các thuật ngữ, trích xuất các khái niệm và trích xuất các quan hệ. Các kỹ thuật thống kê bao gồm giá trị C/NC (C/NC value), phân tích tương phản (contrastive analysis), phân cụm (clustering), phân tích tương quan (co-occurrence analysis), xếp gộp thuật ngữ (term subsumption) và phân cụm phân cấp (ARM).

Kỹ thuật làm giàu ontology miền dựa trên logic

ILP là một môn học học máy xuất phát từ giả thuyết dựa trên kiến thức nền tảng và một tập hợp các ví dụ sử dụng lập trình logic. Trong lĩnh nghiên cứu ontology, ILP được sử dụng ở giai đoạn cuối cùng của các tầng, mức học ontology trong đó các tiên đề tổng quát được thu nhận từ các tiên đề lược đồ (tiên đề với cả ví dụ tích cực và tiêu cực và kiến thức nền tảng).

1.3. ĐÁNH GIÁ KỸ THUẬT LÀM GIÀU ONTOLOGY MIỀN

Đánh giá chất lượng của việc xây dựng ontology là khía cạnh công nghệ web thông minh rất quan trọng vì nó cho phép các nhà nghiên cứu và các nhà chuyên môn đánh giá tính đúng đắn ở mức từ loại, độ bao phủ ở mức khái niệm, tính phù hợp ở mức phân loại và tính đầy đủ ở mức phi phân loại của ontology đã được xây dựng. Đánh giá kỹ thuật làm giàu ontology chia làm 4 nhóm: (1) Đánh giá dựa trên chuẩn vàng; (2) Đánh giá dựa trên khả năng ứng dụng; (3) Đánh giá hướng dữ liệu và (4) Đánh giá con người.

CHƯƠNG 2. XÂY DỰNG MỘT ONTOLOGY MIỀN XUẤT PHÁT

2.1. BÀI TOÁN XÂY DỰNG ONTOLOGY

Natalya F.Noy [4] đã chỉ ra 7 bước chính để xây dựng ontology, bao gồm:

Bước 1: Xác định miền cần xây dựng ontology và phạm vi của việc xây dựng ontology.

Bước 2: Rà soát, phân tích các ontology đã được xây dựng có liên quan đến miền cần xây dựng ontology, qua đó xem xét việc tái sử dụng và tích hợp các ontology đã có.

Bước 3: Phân tích, trích xuất từ các nguồn dữ liệu, qua đó xác định được các khái niệm, thuật ngữ quan trọng của ontology cần xây dựng.

Bước 4: Xác định các khái niệm và cây phân cấp các khái niệm của ontology cần xây dựng.

Bước 5: Định nghĩa các thuộc tính của các khái niệm.

Bước 6: Định nghĩa miền giá trị của các thuộc tính của các khái niệm.

Bước 7: Tạo các thể hiện của các khái niệm và quan hệ giữa các thể hiện của các khái niệm.

2.2. SỰ CẦN THIẾT XÂY DỰNG ONTOLOGY MIỀN XUẤT PHÁT

Natalya F.Noy[4] đã chỉ ra 5 lý do sau đây để xây dựng một ontology: *Thứ nhất*, việc chia sẻ sự 'hiểu' về các cấu trúc thông tin giữa con người và các tác tử phần mềm là mục tiêu lớn nhất trong sự phát triển của ontology. Ví dụ, có rất nhiều Website chứa đựng các thông tin hay dịch vụ về y tế. Nếu các Website này chia sẻ và được xuất bản trên cơ sở sử dụng các thuật ngữ của cùng một ontology thì máy tính có thể trích chọn và tích hợp thông tin từ các nguồn này, trả lời cho các truy vấn người dùng hay là làm input cho một ứng dụng nào khác. *Thứ hai*, việc xây dựng ontology cho phép khả năng sử dụng lại các tri thức miền. *Thứ ba*, việc xây dựng ontology tạo ra các giả thiết tri thức miền rõ ràng. *Thứ tư*, việc xây dựng ontology cho phép tách biệt tri thức miền với tri thức thi hành. *Thứ năm*, phân tích tri thức miền là hoàn toàn có thể thi hành được khi đã biết được các định nghĩa của các khái niệm trong ontology được xây dựng.

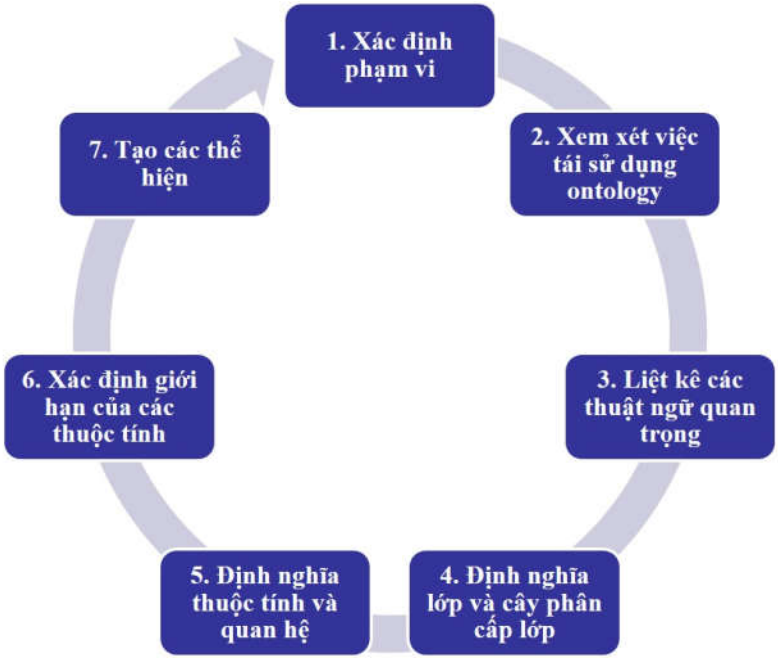
Trong ngành tài nguyên và môi trường, hiện nay có hai bài toán lớn cần giải quyết đó là tích hợp dữ liệu và tìm kiếm ngữ nghĩa. Có nhiều phương pháp, kỹ thuật để giải quyết hai bài toán trên, nhưng phương pháp dựa trên ontology

đang được sử dụng rộng rãi và mang lại hiệu quả cao. Vì vậy, xây dựng ontology cho lĩnh vực tài nguyên và môi trường có vai trò quan trọng trong việc giải quyết các bài toán lớn của ngành. Ngoài ra, ontology được xây dựng cũng là đầu vào quan trọng của các nghiên cứu sâu về các kỹ thuật nâng cấp, làm giàu ontology dựa trên các phương pháp xử lý ngôn ngữ tự nhiên, các phương pháp dựa trên thống kê và các phương pháp dựa trên logic.

2.3. XÂY DỰNG ONTOLOGY MIỀN XUẤT PHÁT CHO MIỀN TÀI NGUYÊN VÀ MÔI TRƯỜNG

2.3.1. Quy trình xây dựng ontology miền xuất phát cho miền tài nguyên và môi trường

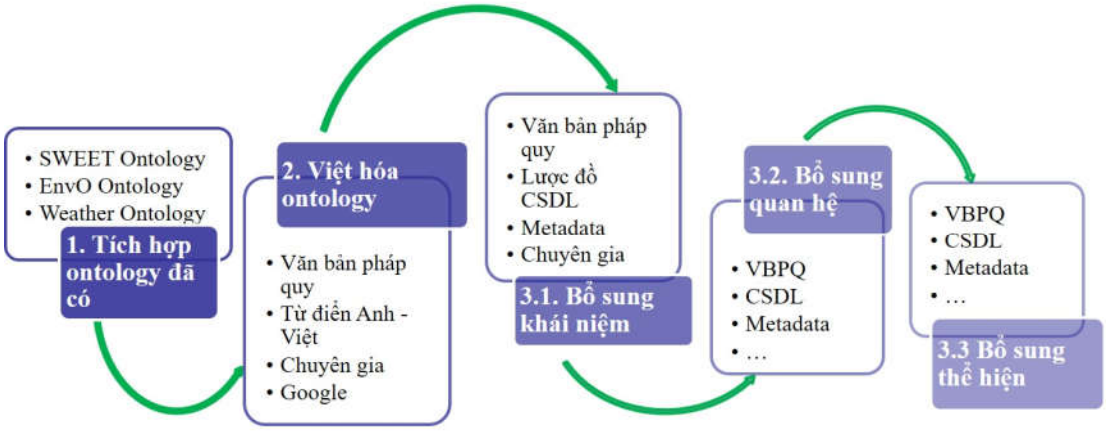
Trên cơ sở phân tích các quy trình xây dựng ontology đang được sử dụng hiện nay trên thế giới và các yếu tố đặc thù về tiếng Việt và các chuyên ngành tài nguyên môi trường, luận án đề xuất quy trình xây dựng ontology tài nguyên môi trường bao gồm 07 bước chính sau:



Hình 2.1. Quy trình xây dựng ontology cho lĩnh vực tài nguyên và môi trường

2.3.2. Phương pháp xây dựng ontology miền xuất phát cho miền tài nguyên và môi trường

Trên cơ sở nghiên cứu các phương pháp, quy trình, giải pháp xây dựng ontology đã có trên thế giới, luận án đề xuất phương pháp khả thi, cụ thể cho việc xây dựng ontology cho lĩnh vực tài nguyên và môi trường bao gồm 03 pha sau:



Hình 2.2. Phương pháp xây dựng ontology cho lĩnh vực tài nguyên và môi trường

Phương pháp xây dựng ontology cho lĩnh vực tài nguyên và môi trường (hình 2.2) bao gồm 03 pha cơ bản sau:

- Pha 1: Tích hợp các ontology đã có liên quan đến ngành tài nguyên và môi trường (trong đó thử nghiệm với 02 lĩnh vực đo đạc bản đồ và khí tượng thủy văn). Mục đích của bước này nhằm tái sử dụng các ontology đã được xây dựng trên thế giới và tại Việt Nam có liên quan đến ngành tài nguyên và môi trường.

- Pha 2: Việt hóa ontology. Với ontology khởi tạo đã được xây dựng trong pha 1 được tích hợp từ các ontology đã có trên thế giới, do đó các khái niệm đa phần là tiếng Anh, nên cần phải chuyển các khái niệm sang tiếng Việt.

- Pha 3: Nâng cấp ontology. Ontology đã được xây dựng từ pha 1 và pha 2

chỉ là ontology khởi tạo, bao gồm số ít các khái niệm và chưa đủ bao quát cho miền tài nguyên và môi trường. Do vậy, cần thiết phải mở rộng, nâng cấp ontology đã có trên cơ sở 03 bước cơ bản: (1) Bổ sung các khái niệm nhằm hiệu chỉnh các khái niệm đã có và mở rộng cây phân cấp khái niệm; (2) Bổ sung các quan hệ giữa các khái niệm và (3) Bổ sung các thể hiện của các khái niệm và các thể hiện của các quan hệ giữa các khái niệm. Nguồn dữ liệu phục vụ trích xuất các khái niệm, quan hệ và các thể hiện là hệ thống các văn bản pháp quy có liên quan, các mô hình dữ liệu, metadata của các CSDL đã được xây dựng trong hệ thống CSDLQg về TN&MT, …

2.3.3. Kết quả xây dựng ontology miền xuất phát cho miền tài nguyên và môi trường

2.3.3.1. Kết quả xây dựng ontology cho lĩnh vực đo đạc và bản đồ

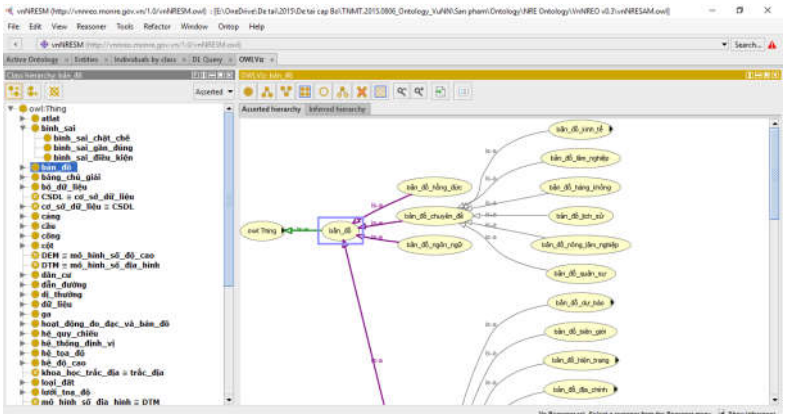
Để xây dựng ontology cho lĩnh vực đo đạc và bản đồ, tác giả sử dụng các nguồn dữ liệu đầu vào như sau:

- Hệ thống văn bản quy phạm pháp luật (khoảng 120 văn bản[4]).
- Từ điển khái niệm, thuật ngữ.
- Cơ sở dữ liệu, quy định kỹ thuật.

Dựa trên ontology ban đầu (được tích hợp sẵn từ ontology tiếng Anh và chuyển sang tiếng Việt), tác giả đã trích xuất các khái niệm từ các văn bản pháp lý và các từ điển, các cơ sở dữ liệu chuyên ngành để bổ sung vào ontology của lĩnh vực đo đạc và bản đồ. Tổng số khái niệm của ontology được xây dựng khoảng 3.000 khái niệm.

Dưới đây là hình ảnh kết quả một số cây phân cấp khái niệm trong ontology cho lĩnh vực đo đạc và bản đồ.

[4]http://vanban.monre.gov.vn/DocViewer.aspx?IDLV=6



Hình 2.3. Cây phân cấp khái niệm "bản đồ"

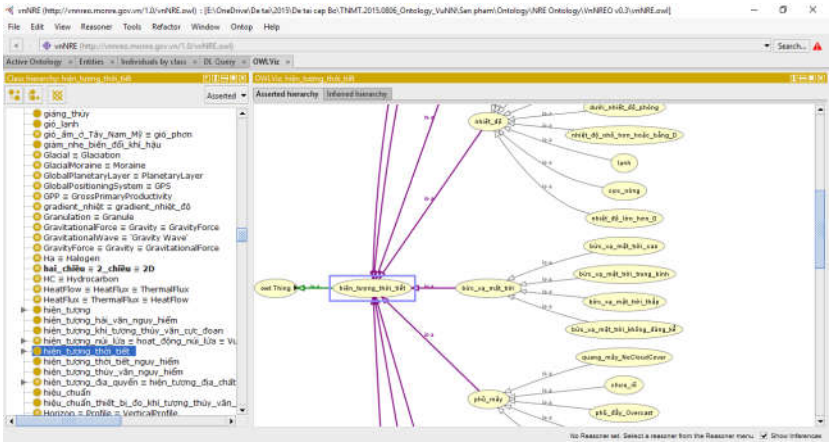
2.3.3.2. Kết quả xây dựng ontology cho lĩnh vực khí tượng thủy văn

Để xây dựng ontology cho lĩnh vực khí tượng thủy văn, tác giả sử dụng các nguồn dữ liệu đầu vào như sau:

- Hệ thống văn bản quy phạm pháp luật (khoảng 100 văn bản[5]).
- Từ điển khái niệm, thuật ngữ.
- Cơ sở dữ liệu, quy định kỹ thuật.

Tổng số khái niệm của ontology được xây dựng khoảng 5.000 khái niệm.

Dưới đây là hình ảnh kết quả một số cây phân cấp khái niệm trong ontology cho lĩnh vực khí tượng thủy văn.



Hình 2.4. Cây phân cấp khái niệm "hiện tượng thời tiết"

2.3.3.3. Kết quả xây dựng ontology cho miền tài nguyên và môi trường

[5]http://vanban.monre.gov.vn/DocViewer.aspx?IDLV=5

Trên cơ sở ontology đã xây dựng, tác giả đã xây dựng bộ từ điển khái niệm thuật ngữ bao gồm 111.150 khái niệm (trong đó 20.055 khái niệm có song ngữ Anh - Việt, 27.322 khái niệm có quan hệ đồng nghĩa). Bộ từ điển khái niệm này có số lượng khái niệm, thuật ngữ rất lớn (Mạng từ tiếng Việt viet.wordnet.vn bao gồm 67.344 khái niệm) rất có giá trị phục vụ các nghiên cứu về xử lý ngôn ngữ tự nhiên, phân tích và khai phá dữ liệu, trích rút thông tin, …và chia sẻ cho cộng đồng khai thác sử dụng.

CHƯƠNG 3. MỘT MÔ HÌNH HỌC SÂU SUỐT ĐỜI LÀM GIÀU THỂ HIỆN ONTOLOGY MIỀN

3.1. HỌC SUỐT ĐỜI VÀ BÀI TOÁN NHẬN DẠNG THỰC THỂ

Học suốt đời

Học máy suốt đời (Lifelong Machine Learning: LML) là một quá trình học liên tục. Tại thời điểm bất kỳ, bộ học đã thực hiện một chuỗi N bài toán học, $\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_N$. Các bài toán này, còn được gọi là các bài toán trước (previous tasks) có các tập dữ liệu tương ứng là $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_N$. Các bài toán có thể cùng kiểu hoặc thuộc các kiểu khác nhau và từ cùng một miền ứng dụng hoặc các miền ứng dụng khác nhau. Khi đối mặt với bài toán thứ N+1, $\mathcal{T}_{N+1}$ (được gọi là bài toán mới hoặc bài toán hiện tại) với dữ liệu $\mathcal{D}_{N+1}$, bộ học có thể tận dụng tri thức quá khứ trong cơ sở tri thức (KB) để hỗ trợ học bài toán $\mathcal{T}_{N+1}$.

Mục tiêu của LML thường là tối ưu hóa hiệu năng của bài toán mới $\mathcal{T}_{N+1}$, song nó có thể tối ưu hóa bất kỳ bài toán nào bằng cách xử lý các bài toán còn lại như các bài toán trước đó. Việc cập nhật tri thức có thể bao gồm liên quan đến kiểm tra tính nhất quán, lập luận và biến đổi của tri thức mức cao bổ sung vào KB.

LML có 3 đặc điểm chính: (1) Quá trình học liên tục, (2) Tích luỹ và lưu giữ tri thức trong cơ sở tri thức (KB), (3) Khả năng sử dụng các tri thức đã học trước đó để xử lý các bài toán mới.

Trường điều kiện ngẫu nhiên (CRF)

Trường điều kiện ngẫu nhiên (Conditional Random Field - CRF) được giới thiệu vào những năm 2001 bởi Lafferty và các đồng nghiệp **Error! Reference source not found.**. CRF là một nền tảng để xây dựng mô hình xác suất để phân đoạn và gán nhãn chuỗi. Trường điều kiện ngẫu nhiên dựa trên ý tưởng gốc từ mô hình Markov ẩn (Hidden Markov Model) và được cải thiện để khắc phục các nhược điểm của nó cũng như của mô hình markov entropy cực đại (Maximum Entropy Markov Model, MEMM).

Bộ nhớ dài ngắn hai chiều (Bi-LSTM)

Bộ nhớ dài ngắn hai chiều (LSTM) được biết đến như là một biến thể của

mạng nơron tích chập (RNN), ban đầu được đưa ra như là một giải pháp để giải quyết vấn đề lãng quên tri thức trong mạng nơron (vanishing and exploding gradient) và do đó cho phép các mạng sâu thực thi tốt hơn trong thực tế (S. Hochreiter and J. Schmidhuber[5]). Ý tưởng này đã được thực hiện trong các LSTM cell bằng cách tạo ra một trạng thái nhớ bên trong, trong đó đơn giản là bổ sung vào đầu vào đã được xử lý để giảm ảnh hưởng nhiều lần của các giá trị gradient nhỏ.

Mô hình kết hợp Bi-LSTM và CRF

Trong mô hình Bi-LSTM, nhãn cuối cùng của đầu ra không được xác định bởi hàm softmax. Điều đó có nghĩa là việc gán nhãn cuối cùng cho một từ không phụ thuộc vào nhãn của các từ xung quanh nó. Vì vậy, với ưu điểm của cho phép gán nhán theo ngữ cảnh trong CRF [6], việc bổ sung một lớp CRF vào mô hình LSTM hay mô hình Bi-LSTM sẽ cho phép mô hình này học việc gán nhãn chuỗi một cách tốt nhất (gọi là mô hình kết hợp Bi-LSTM+CRF), nên có thể tối đa hóa độ chính xác của mô hình.

3.2. MÔ HÌNH HỌC SÂU SUỐT ĐỜI MỨC KÝ TỰ CHO NHẬN DẠNG THỰC THỂ TRONG VĂN BẢN TIẾNG VIỆT

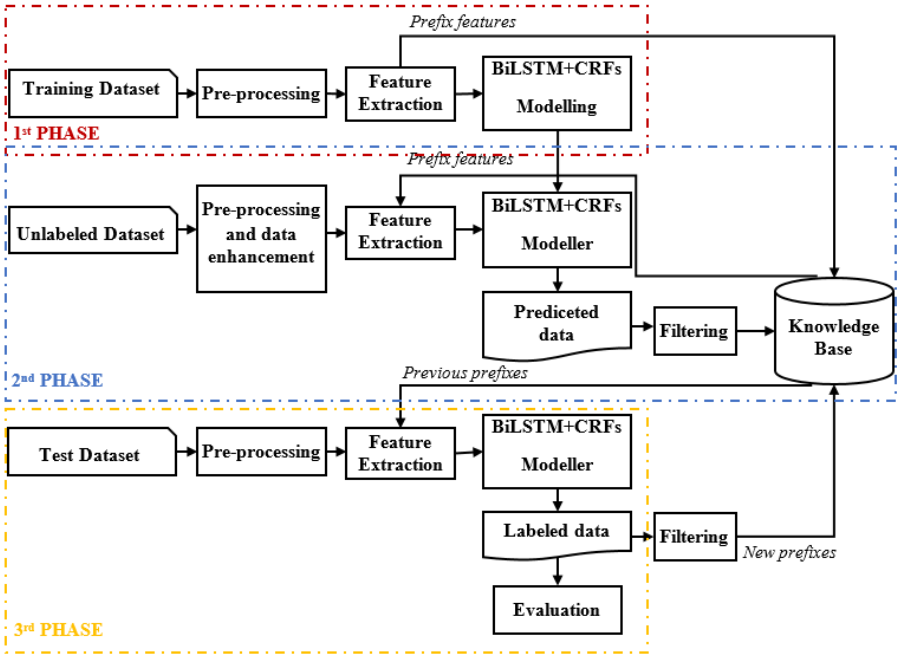
Tiền tố là các từ hay tập hợp từ thường có vị trí ở trước của các thực thể dạng tên trong câu. Ví dụ: tiền tố "Công ty" thường đứng trước tên của một tổ chức (nhãn ORG); tiền tố "Ông", "Bà" thường đứng trước các thực thể tên người (nhãn PER), …Danh sách các tiền tố có thể mở rộng qua các lần học, nên đây là đặc trưng quan trọng để có thể sử dụng trong học suốt đời.

Giả sử K là tập tiền tố tin cây được trích xuất từ các công việc trước đó sử dụng mô hình gọi là M, trong đó mô hình kết hợp Bi-LSTM+CRF đã được sử dụng. Mô hình M được huấn luyện dựa trên tập dữ liệu huấn luyện D^t. Ban đầu, tập K chính là tập K^t (tập tất cả các tiền tố tin cậy của tập dữ liệu huấn luyện D^t). Giả sử M xử lý nhiều bài toán hơn và nhiều tiền tố tin cậy được trích xuất, theo đó kích thước tập K cũng sẽ lớn hơn. Khi xử lý bài toán D_{n+1}, tập K cho phép trích xuất đặc trưng tiền tố được nhiều hơn, mô hình M có thể cho kết quả tốt hơn đối với bài toán mới. Mô hình được đề xuất trong Hình 3.1 gồm 3 pha chính:

- Huấn luyện mô hình

- Trích xuất đặc trưng suốt đời

- Đánh giá mô hình đề xuất



Hình 3.1. DeepLML-NER: Mô hình học sâu suốt đời mức ký tự cho nhận dạng thực thể tiếng Việt[NNVu5]

3.2.1. Tinh chỉnh nhằm nâng cao chất lượng dữ liệu huấn luyện

Dữ liệu được thu thập từ trang tin tức tiếng Việt sau đó chúng tôi sử dụng công cụ để phân đoạn thành các câu tiếng Việt. Tuy nhiên, chúng tôi phát hiện ra có một số lỗi như sau: (1) Có nhiều câu quá ngắn, ví dụ như tiêu đề của các bài báo hoặc mô tả của các ảnh trong bài báo.(2) Các câu quá dài, nguyên nhân là do lỗi của công cụ phân đoạn câu, trong đó có lỗi trong việc phân đoạn 2 đến 3 câu liên tiếp.(3) Một âm tiết (từ đơn) trong tiếng Việt có không quá 7 ký tự (ví dụ: từ đơn dài nhất là từ "nghiêng" có 7 ký tự).(4) Các từ nghép tiếng Việt thông thường được tạo thành từ 2 từ đơn. Do vậy, tác giả đã xây dựng công cụ tiền xử lý dữ liệu, lọc bỏ các dữ liệu gặp các lỗi trên, sau khi lọc bỏ các câu bị lỗi, tổng số câu của bộ dữ liệu dantri giảm khoảng 15%.

3.2.2. Tối ưu hóa các tham số mô hình

Theo mô hình Bi-LSTM-CRF do Pham và công sự [13] đề xuất trong mã nguồn đã được xuất bản trên Github6, lần chạy thực nghiệm đầu tiên chúng tôi đã sử dụng các tham số mặc định như Pham đã sử dụng. Trong quá trình thực nghiệm tiếp theo, chúng tôi đã thử các tổ hợp giá trị các tham số khác nhau sau đó đã chọn ra được các giá trị tham số tối ưu với bài toán trong nghiên cứu của chúng tôi. Bảng 1 liệt kê giá trị các tham số đã được điều chỉnh và được sử dụng trong thực nghiệm của nghiên cứu trong chuyên đề này.

Tham số	Giá trị mặc định	Giá trị điều chỉnh
Số đơn vị trong 2 lớp LSTM (word_lstm_units)	100	200
Số chiều đặc trưng tiền tố (pre_word_feature_size)	100	200
Kích thước lô (batch_size)	40	64
Tốc độ học (learning_rate)	0.01	0.001

Bảng 3.1. Giá trị các tham số điều chỉnh của mô hình

3.2.3. Trích xuất đặc trưng suốt đời

Thuật toán trích xuất đặc trưng suốt đời:

1 $K_p \leftarrow \emptyset$

2 **loop:**

3 $F \leftarrow FeatureGeneration\ (D_{n+1}, K)$

4 $E_{n+1} \leftarrow Apply - Model(M, F)$

5 $S \leftarrow S \cup \{E_{n+1}\}$

6 $K_{n+1} \leftarrow Frequent - prefixs - Mining(S, \lambda)$

7 **if** $K_p = K_{n+1}$ **then:**

8 **break**

9 **else:**

10 $K \leftarrow K^t \cup K_{n+1}$

11 $K_p \leftarrow K_{n+1}$

[6] https://github.com/pth1993/NNVLP

12 $S \leftarrow S - \{E_{n+1}\}$

13 **end if**

14 **end loop**

Giải thích các bước trong thuật toán:

1. Sinh đặc trưng F trên tập dữ liệu D_{n+1} và áp dụng vào mô hình M để sinh ra tập các thực thể E_{n+1} (dòng 3)

2. E_{n+1} (kết quả thu được khi sử dụng mô hình M) được thêm vào tập S - kho thông tin quá khứ. Từ S, khai phá ra các tiền tố thường xuyên K_{n+1} sử dụng ngưỡng λ.

3. Nếu tập K_{n+1} giống với tập K_p từ vòng lặp trước, có nghĩa là không có tiền tố nào được tìm thấy thì vòng lặp sẽ dừng lại.

4. Nếu không, có nghĩa rằng có các tiền tố tin cậy mới được tìm thấy. M có thể gán nhãn chính xác hơn trong vòng lặp tiếp theo. Dòng 10 và 11 cập nhật lại hai tập dữ liệu cho vòng lặp sau.

3.3. THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.3.1. Dữ liệu

Dữ liệu VLSP 2018

Thử nghiệm được thực hiện với bộ dữ liệu cho nhận dạng thực thể được cung cấp trong khuôn khổ cuộc thi của VLSP 2018. Tập dữ liệu này được cung cấp bởi cộng đồng xử lý ngôn ngữ tiếng Việt, được thu thập từ các trang báo điện tử tiếng Việt cho 10 miền. Dưới đây là bảng thống kê cho mỗi tập dữ liệu của bộ dữ liệu VLSP 2018.

Tập dữ liệu	**PER**	**ORG**	**LOC**	**MICS**	**Total**
Training	4,600	5,587	6,289	743	17,219
Developme nt	492	723	795	63	2,073
Test	1,883	2,126	2,377	178	6,564
Tổng	**6,978**	**8,436**	**9,461**	**984**	**25,856**

Bảng 3.2. Thống kê bộ dữ liệu VLSP 2018

Tập dữ liệu được chia thành 3 phần Train, Dev và Test với 10 tập dữ liệu nhỏ hơn theo các lĩnh vực. **Error! Reference source not found.** thống kê số lượng thực thể chia theo từng miền của tập dữ liệu VLSP 2018.

Miền	Đời sống	Giải trí	Giáo dục	KHCN	Kinh tế	Pháp luật	Thế giới	Thể thao	Văn hóa	Xã hội
Đời sống	-	0.14	0.13	0.1	0.14	0.14	0.09	0.11	0.15	0.13
Giải trí	0.05	-	0.05	0.05	0.07	0.04	0.04	0.06	0.08	0.05
Giáo dục	0.06	0.07	-	0.06	0.12	0.11	0.05	0.06	0.12	0.12
KHCN	0.06	0.08	0.07	-	0.13	0.06	0.12	0.08	0.12	0.1
Kinh tế	0.04	0.06	0.08	0.07	-	0.08	0.06	0.05	0.09	0.11
Pháp luật	0.05	0.04	0.08	0.04	0.09	-	0.04	0.03	0.08	0.1
Thế giới	0.03	0.05	0.04	0.08	0.08	0.04	-	0.05	0.09	0.07
Thể thao	0.03	0.05	0.04	0.04	0.05	0.03	0.04	-	0.05	0.04
Văn hóa	0.03	0.05	0.06	0.05	0.07	0.06	0.05	0.04	-	0.08
Xã hội	0.04	0.05	0.08	0.05	0.11	0.08	0.05	0.04	0.1	-

Bảng 3.3. So sánh số thực thể giao nhau giữa các miền trong tập dữ liệu VLSP2018

Dữ liệu dantri

Tập dữ liệu thứ hai được sử dụng trong thử nghiệm là tập dữ liệu chưa gán nhãn, được thu thập từ 1.600 bài báo thuộc 16 miền từ trang tin tức tiếng Việt[7]. Bộ dữ liệu này chứa 246.586 câu bao gồm 6.682.201 từ. Thống kê chi tiết của bộ dữ liệu dantri được mô tả trong bảng 3.4.

Miền	Số từ	Số câu
Chuyện lạ	350,450	15,169
Giải trí	271,666	11,086
Giáo dục	680,809	25,331

[7] http://dantri.com.vn

Kinh doanh	483,219	15,795
Nhịp sống trẻ	309,252	10,910
Ô tô - Xe máy	480,321	14,680
Pháp luật	462,295	16,003
Sức khỏe	475,327	17,885
Sức mạnh	427,959	14,374
Sự kiện	404,959	14,480
Tấm lòng nhân ái	180,972	6,746
Thế giới	401,711	14,664
Thể thao	402,051	17,215
Tình yêu giới tính	514,916	23,872
Văn hóa	433,822	15,947
Xã hội	402,472	13,463
Tổng	**6,682,201**	**246,586**

Bảng 3.4. Thống kê bộ dữ liệu dantri

3.3.2. Thiết lập thử nghiệm

Thiết lập tham số:Chúng tôi đã thiết lập số chiều nhúng từ và số chiều đặc trưng tiền tố là 100. Do tập dữ liệu huấn luyện rất lớn, do đó chúng tôi chọn kích thước lô là 40 và số đơn vị trong 2 lớp LSTM là 100. Số bộ lọc CNN là 30,và kích thước cửa sổ CNN là 3. Chúng tôi thiết lập các tham số của bộ thư viện theo mặc định: ví dụ, $learningrate = 0{:}01$; $\beta_1 = 0{:}9$; $\beta_2 = 0{:}999$; $\epsilon = 10^{-7}$. Giá trị ngưỡng λđể lọc các tiền tố tin cậy là 2.

3.3.3. Kết quả thử nghiệm và phân tích

Chúng tôi đã thực hiện các thử nghiệm theo các kịch bản khác nhau qua đó để đánh giá hiệu quả của các đề xuất của chúng tôi trong nghiên cứu này.

Kết quả thử nghiệm chỉ với mô hình Deep LML

Kịch bản thử nghiệm này nhằm đánh giá mô hình Deep LML so với các phương pháp cơ sở (BiLSTM+CRF và CRFs). Bảng 3.5 chỉ ra các kết quả thử nghiệm chi tiết của mô hình đã đề xuất.

Miền	CRF			Bi-LSTM+CRF			Deep LML (1 lần chạy)		
	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
Đời sống	75.71	67.05	70.36	65.82	67.89	66.84	52.33	71.10	60.28
Giải trí	64.00	53.96	55.73	63.86	70.00	66.79	70.72	72.62	71.65
Giáo dục	70.83	63.42	66.27	78.36	76.14	77.23	82.44	83.92	83.17
KHCN	60.18	62.80	57.89	66.47	53.74	59.43	51.33	45.54	48.26

Kinh tế	74.38	64.53	67.12	69.63	69.38	69.50	72.75	69.89	71.29
Pháp luật	83.47	75.78	78.92	83.73	79.05	81.32	90.96	85.77	88.29
Thế giới	50.00	56.55	59.08	67.02	62.20	64.52	71.53	67.91	69.67
Thể thao	62.62	37.47	42.54	39.39	49.79	43.99	56.62	62.45	59.39
Văn hóa	62.53	49.55	53.17	61.18	62.75	61.95	61.38	61.38	61.38
Xã hội	82.38	68.57	74.23	75.47	73.97	74.71	65.32	64.62	64.97
Trung bình	68.61	59.97	62.53	67.09	66.49	66.63	**67.54**	**68.52**	**67.84**

Bảng 3.5. Kết quả thực nghiệm chỉ với mô hình Deep LML sau 1 lần chạy

Bởi vì mô hình Deep LML có đặc tính ngẫu nhiên, do vậy với mỗi lần chạy sẽ cho các kết quả khác nhau. Chúng tôi thực hiện chạy thực nghiệm với 10 lần, sau đó chọn kết quả tốt nhất, kết quả được chỉ ra trong Bảng 3.6.

Miền	**Deep LML (1 lần chạy)**			**Deep LML (tốt nhất trong 10 lần chạy)**		
	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
Đời sống	52.33	71.10	60.28	70.41	77.50	73.80
Giải trí	70.72	72.62	71.65	68.38	68.68	68.53
Giáo dục	82.44	83.92	83.17	80.12	75.69	77.84
KHCN	51.33	45.54	48.26	67.63	53.92	60.00
Kinh tế	72.75	69.89	71.29	68.76	72.88	70.76
Pháp luật	90.96	85.77	88.29	84.91	87.10	85.99
Thế giới	71.53	67.91	69.67	66.11	63.41	64.73
Thể thao	56.62	62.45	59.39	43.93	61.43	51.23
Văn hóa	61.38	61.38	61.38	61.95	63.79	62.85
Xã hội	65.32	64.62	64.97	77.02	80.52	78.73
Trung bình	67.54	68.52	67.84	**68.92**	**70.49**	**69.45**

Bảng 3.6. Kết quả thực nghiệm chỉ với mô hình Deep LML tốt nhất sau 10 lần chạy

Kết quả thử nghiệm mô hình Deep LML với dữ liệu đã được tinh chỉnh

Sau các bước nâng cao chất lượng dữ liệu (đã mô tả trong mục 3.2.2), số lượng các câu trong tập dữ liệu của mỗi miền trung bình giảm từ 10% đến 20%. Trong thử nghiệm này, chúng tôi đã sử dụng cùng các giá trị các tham số như thử nghiệm trong phần 3.3.3.1 với các thiết lập khởi tạo ban đầu. Bảng 3.8 chỉ ra các kết quả thử nghiệm và chứng tỏ ảnh hưởng của việc tinh chỉnh dữ liệu đến kết quả cuối cùng.

Miền	**Deep LML (1 lần**	**Deep LML (tốt nhất**	**Deep LML (với dữ liệu**

	chạy)			trong 10 lần chạy)			đã được tinh chỉnh)		
	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
Đời sống	52.33	71.10	60.28	70.41	77.50	73.80	54,92	76,11	63.80
Giải trí	70.72	72.62	71.65	68.38	68.68	68.53	68,22	76,81	72.26
Giáo dục	82.44	83.92	83.17	80.12	75.69	77.84	79,84	93,52	86.14
KHCN	51.33	45.54	48.26	67.63	53.92	60.00	51,97	47,57	49.68
Kinh tế	72.75	69.89	71.29	68.76	72.88	70.76	74,06	72,73	73.39
Pháp luật	90.96	85.77	88.29	84.91	87.10	85.99	90,96	84,13	87.41
Thế giới	71.53	67.91	69.67	66.11	63.41	64.73	74,50	70,33	72.36
Thể thao	56.62	62.45	59.39	43.93	61.43	51.23	61,49	63,83	62.64
Văn hóa	61.38	61.38	61.38	61.95	63.79	62.85	67,27	66,11	66.68
Xã hội	65.32	64.62	64.97	77.02	80.52	78.73	64,29	65,41	64.85
Trung bình	67.54	68.52	67.84	68.92	70.49	69.45	**68,75**	**71,65**	**69.92**

Bảng 3.7. Kết quả mô hình Deep LML với dữ liệu đã tinh chỉnh

Kết quả thử nghiệm mô hình Deep LML với dữ liệu đã được tinh chỉnh và tối ưu tham số

Trong kịch bản thử nghiệm này, chúng tôi sử dụng cả dữ liệu đã được tinh chỉnh (mô tả trong mục 3.2.2) và các tham số tối ưu (mục 3.2.3, trong bảng 3.1). Kết quả được đưa ra trong bảng 3.9.

Miền	Deep LML (1 lần chạy)			Deep LML với dữ liệu đã tinh chỉnh			Deep LML với dữ liệu đã tinh chỉnh và tối ưu tham số		
	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
Đời sống	52.33	71.10	60.28	54,92	76,11	63.80	60.65	70.88	65.37
Giải trí	70.72	72.62	71.65	68,22	76,81	72.26	73.15	78.29	75.63
Giáo dục	82.44	83.92	83.17	79,84	93,52	86.14	83.74	86.26	84.98
KHCN	51.33	45.54	48.26	51,97	47,57	49.68	53.26	49.27	51.19
Kinh tế	72.75	69.89	71.29	74,06	72,73	73.39	72.23	75.25	73.71
Pháp luật	90.96	85.77	88.29	90,96	84,13	87.41	89.65	90.59	90.12
Thế giới	71.53	67.91	69.67	74,50	70,33	72.36	72.31	69.25	70.75
Thể thao	56.62	62.45	59.39	61,49	63,83	62.64	58.76	63.86	61.21
Văn hóa	61.38	61.38	61.38	67,27	66,11	66.68	68.01	67.11	67.56
Xã hội	65.32	64.62	64.97	64,29	65,41	64.85	64.29	65.41	64.85
Trung bình	67.54	68.52	67.84	68,75	71,65	69.92	**69.60**	**71.62**	**70.53**

Bảng 3.8. Kết quả mô hình Deep LML với dữ liệu đã tinh chỉnh và tối ưu tham số

So sánh với kết quả của kịch bản thử nghiệm trong mục 3.3.3.1 (1 lần chạy), việc kết hợp chạy mô hình Deep LML với dữ liệu sau khi tinh chỉnh và các tham số tối ưu, kết quả tốt hơn và tăng trung bình 2.67%. Kết quả này là bằng chứng cho hướng đi đúng đắn của chúng tôi.

CHƯƠNG 4. MỘT MÔ HÌNH SINH TAXONOMY CHO ONTOLOGY MIỀN

4.1. GIỚI THIỆU CHUNG VỀ SINH TAXONOMY

4.1.1. Khái niệm sinh taxonomy

Taxonomy là nguồn ngữ nghĩa giúp phân loại và bổ sung ngữ nghĩa cho dữ liệu. Sinh taxonomy đề cập đến quá trình tạo ra các chủ đề hoặc các khái niệm và các quan hệ của chúng từ kho văn bản đầu vào. Công việc sinh taxonomy về cơ bản là liệt kê các chủ đề hoặc các danh mục từ kho văn bản và liên kết từng chủ đề với các văn bản có liên quan

4.1.2. Sinh taxonomy và các công việc liên quan

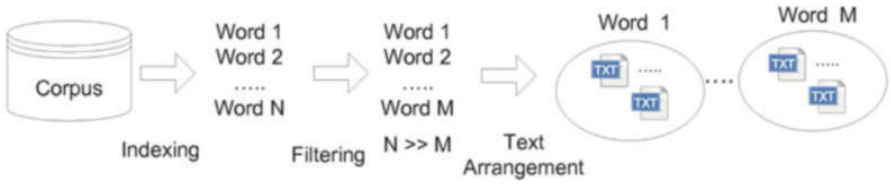
Các công việc liên quan đến sinh taxonomy bao gồm: (i) trích xuất từ khóa và các mối liên quan của nó đến sinh văn bản; (ii) phân loại từ là việc một từ đơn (single word) hay bigram (n-gram với n=2) được phân loại thành một hoặc một vài chủ đề được định nghĩa trước (thay vì phân loại văn bản); (iii) phân cụm từ (liên quan đến phân lớp từ) trong đó các từ được phân cụm theo ngữ nghĩa vào các cụm; (iv) định tuyến chủ đề (topic routing) là nhiệm vụ ngược lại với phân loại văn bản, trong đó một chủ đề được đưa ra làm đầu vào và đầu ra là danh sách các văn bản thuộc về chủ đề đó.

4.2. CÁC PHƯƠNG PHÁP SINH TAXONOMY

Sinh taxonomy từ kho văn bản, được chia thành các nhóm phương pháp: (i) phương pháp dựa trên chỉ mục; (ii) phương pháp dựa trên phân cụm; (iii) phương pháp dựa trên kết hợp và (iv) phương pháp dựa trên phân tích liên kết.

4.2.1. Sinh taxonomy dựa trên chỉ mục

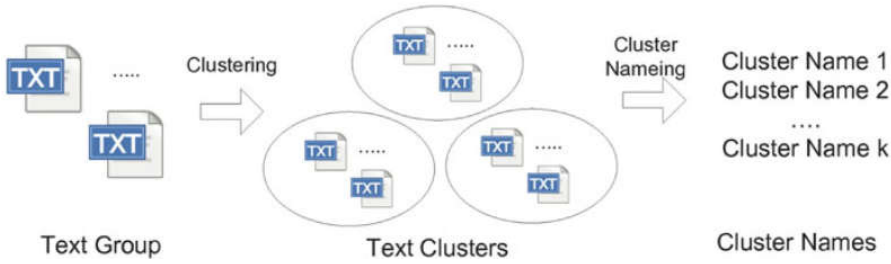
Hình 4.11 minh họa quá trình sinh taxonomy dựa trên lập chỉ mục văn bản. Đầu vào của quá trình sinh taxonomy là một kho văn bản. Nó được lập chỉ mục thành một danh sách các từ, một số trong số chúng được chọn làm taxonomy và các văn bản có liên quan được sắp xếp theo mỗi taxonomy. Các từ được chọn là danh mục hoặc chủ đề của một nhóm văn bản đã cho.



Hình 4.1. Quá trình sinh taxonomy dựa trên lập chỉ mục văn bản[7]

4.2.2. Sinh taxonomy dựa trên phân cụm

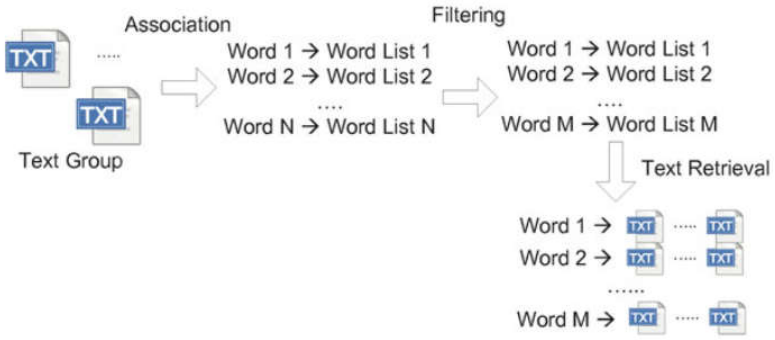
Phần này trình bày quá trình sinh taxonomy dựa trên các phương pháp phân cụm văn bản (hình 4.2). Đầu vào là một kho văn bản và các văn bản trong kho văn bản được nhóm lại thành các nhóm nhỏ. Mỗi cụm được đặt tên theo quy trình và các cụm được đặt tên này được tạo ra thông qua quá trình phân cụm văn bản và các cụm đã được đặt tên này chính là kết quả đầu ra của quá trình sinh taxonomy.



Hình 4.2. Phương pháp sinh taxonomy dựa trên phân cụm[7]

4.2.3. Sinh taxonomy dựa trên luật kết hợp

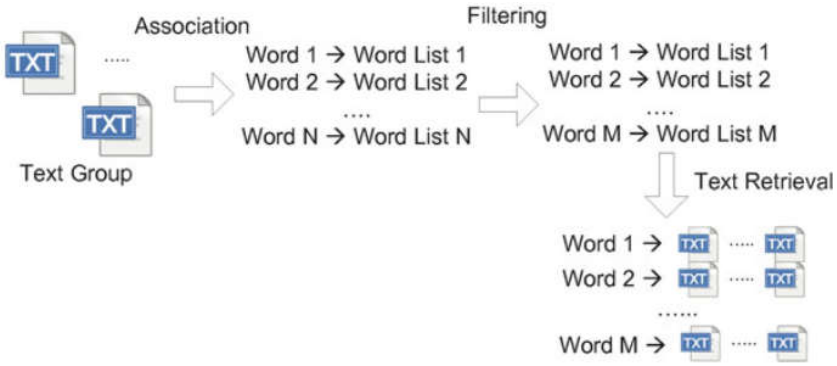
Hình 4.3 mô tả phương pháp sinh taxonomy dựa trên kết hợp từ. Đầu vào là một kho văn bản và các văn bản riêng lẻ trong kho văn bản đó được lập chỉ mục thành một tập hợp các từ. Các luật kết hợp được trích xuất từ tập các từ. Một vài luật kết hợp được lọc và các từ trong các phần có điều kiện được đưa ra như danh sách của các taxonomy.



Hình 4.3. Phương pháp dựa trên luật kết hợp[7]

4.2.4. Sinh taxomony dựa trên phân tích kết nối

Trong các mục 4.2.1, 4.2.2 và 4.2.3 đã trình bày ba phương pháp sinh taxonomy dựa trên lập chỉ mục, dựa trên phân cụm và dựa trên luật kết hợp. Chúng ta đã định nghĩa các kết nối giữa các văn bản trong kho văn bản như là một mạng và việc chọn lựa các văn bản đóng vai trò trung tâm trọng mạng. Hình 4.14 đã chỉ rõ, các taxonomy được sinh ra bằng cách lập chỉ mục các văn bản đã chọn được gọi là các văn bản trung tâm. Phần này sẽ trình bày chi tiết về phương pháp sinh taxonomy trong hình 4.4.



Hình 4.4. Phương pháp dựa trên phân tích kết nối[7]

4.3. MÔ HÌNH SINH TAXONOMY CHO ONTOLOGY MIỀN TÀI NGUYÊN VÀ MÔI TRƯỜNG

Mô hình sinh taxonomy cho ontology miền tài nguyên và môi trường bao gồm công việc chính:

(i) Trích chọn thuật ngữ, khái niệm từ kho dữ liệu văn bản pháp luật tài nguyên môi trường

Kết quả của công việc (i) là danh sách các thuật ngữ, khái niệm được trích xuất từ kho văn bản pháp luật tài nguyên và môi trường

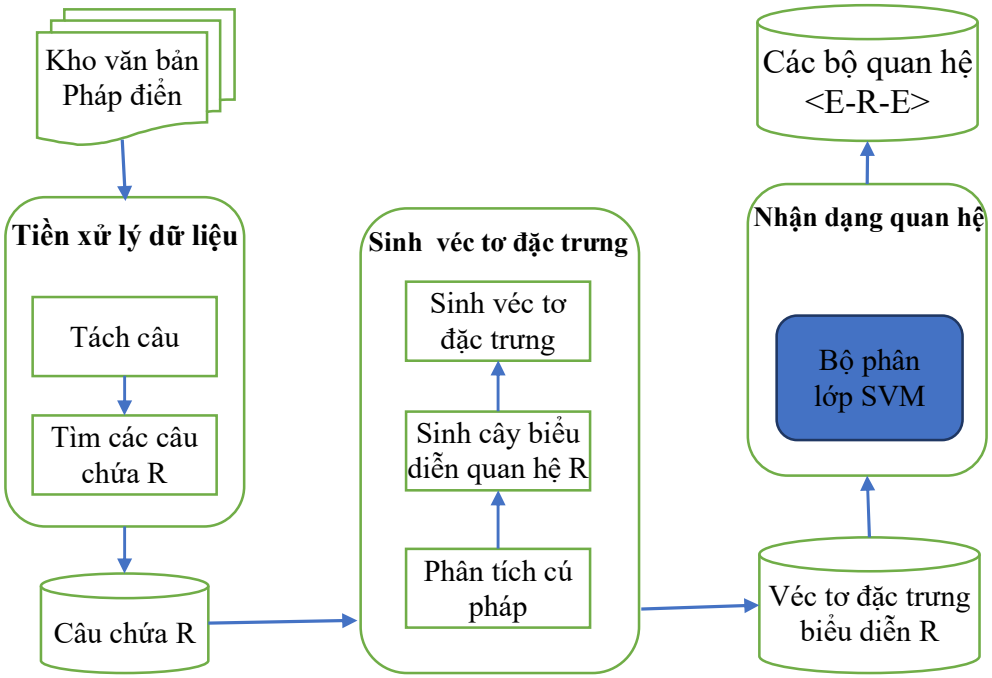
(ii) Trích chọn quan hệ giữa các thuật ngữ từ kho dữ liệu văn bản pháp luật tài nguyên và môi trường, được chia làm ba pha bao gồm:

- Tiền xử lý dữ liệu;

- Sinh véc tơ đặc trưng;

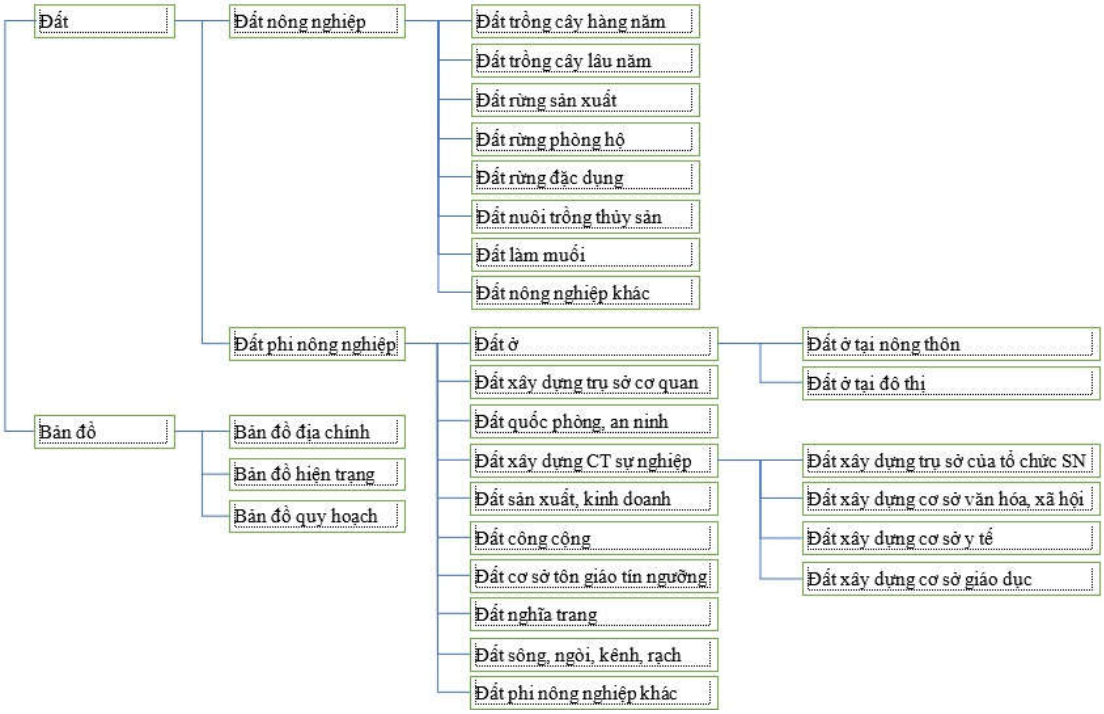
- Nhận dạng quan hệ.

Mô hình trích chọn quan hệ gồm 3 pha: (i) Tiền xử lý dữ liệu; (ii) sinh véc tơ đặc trưng và (iii) nhận dạng quan hệ, được mô tả trong hình 4.5:



Hình 4.5. Mô hình trích chọn quan hệ dựa trên cây phân tích cú pháp

Kết quả thực nghiệm với bộ dữ liệu đầu vào là pháp điển cho chủ đề đất đai như sau (hình 4.6):

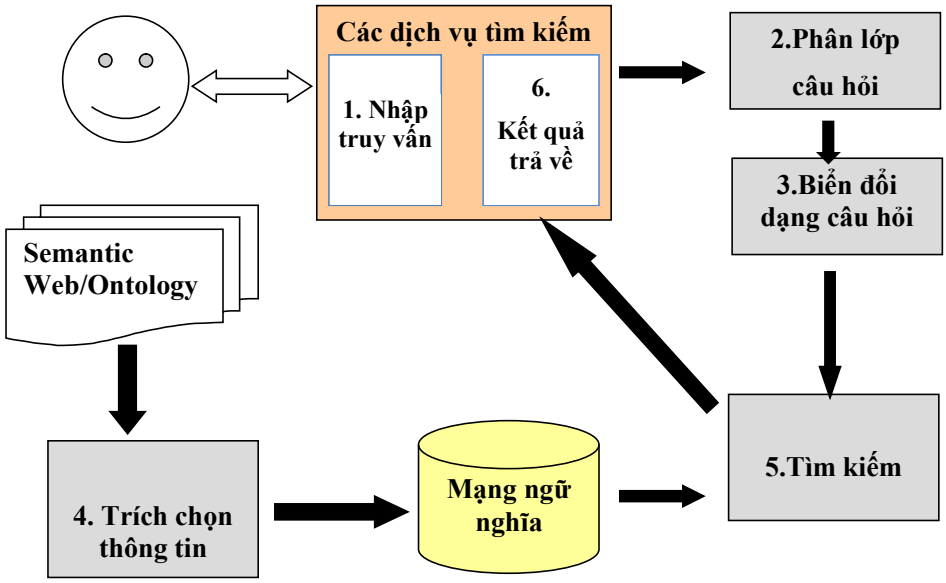


Hình 4.6. Ví dụ taxonomy được sinh với dữ liệu đầu vào là pháp điển cho chủ đề đất đai

CHƯƠNG 5. MỘT MÔ HÌNH TÌM KIẾM NGỮ NGHĨA DỰA TRÊN ONTOLOGY MIỀN

5.1. GIỚI THIỆU CHUNG VỀ TÌM KIẾM NGỮ NGHĨA

Mô hình kiến trúc một máy tìm kiếm ngữ nghĩa được mô tả như hình 4.2

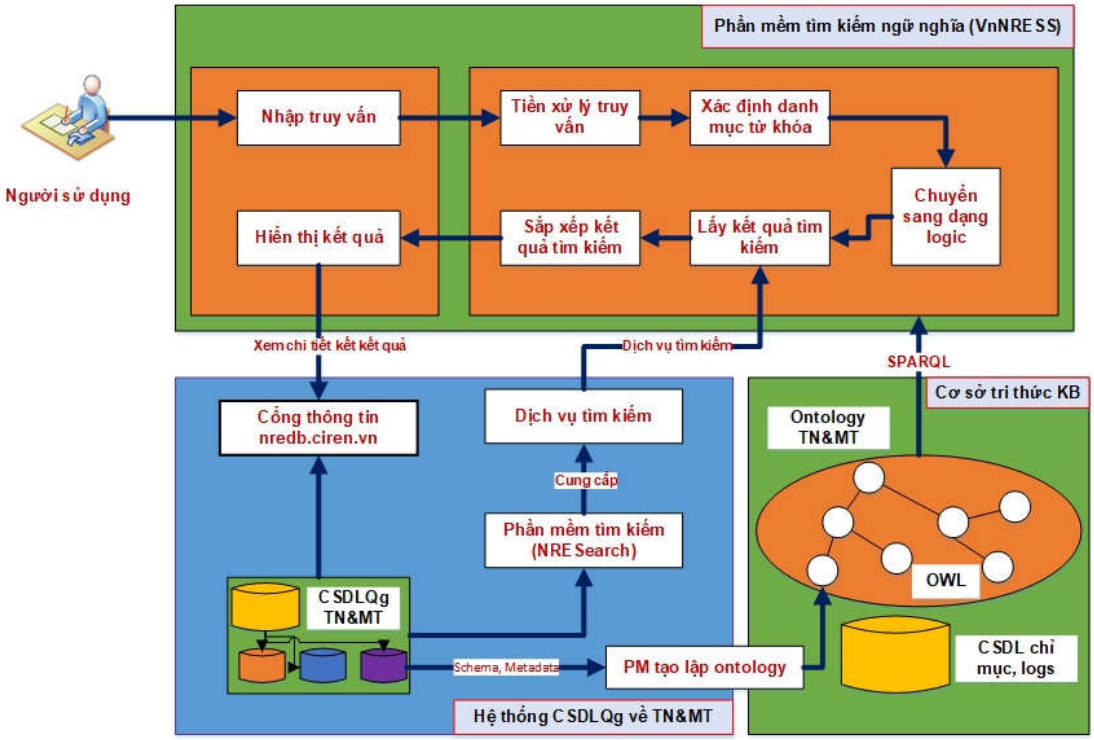


Hình 5.1. Kiến trúc một máy tìm kiếm ngữ nghĩa

Có thể thấy rằng sự khác biệt trong cấu trúc của máy tìm kiếm ngữ nghĩa so với máy tìm kiếm thông thường nằm ở phần kiến trúc bên trong, cụ thể ở hai thành phần: phân tích câu hỏi và tập dữ liệu tìm kiếm.

5.2. MÔ HÌNH TÌM KIẾM NGỮ NGHĨA DỰA TRÊN ONTOLOGY MIỀN

Trên cơ sở mô hình kiến trúc chung cho hệ thống tìm kiếm ngữ nghĩa và các nội dung khảo sát, phân tích của hệ thống CSDLQg về TNMT, nhóm tác giả đề xuất mô hình tìm kiếm ngữ nghĩa dựa trên ontology cho CSDLQg về TNMT như sau:



Hình 5.2. Mô hình kiến trúc đề xuất cho hệ thống tìm kiếm ngữ nghĩa của CSDLQg về TNMT

5.3. XÂY DỰNG PHẦN MỀM TÌM KIẾM NGỮ NGHĨA DỰA TRÊN ONTOLOGY CHO CSDLQG VỀ TNMT

Ontology

Ontology được sử dụng cho thử nghiệm hệ thống tìm kiếm ngữ nghĩa dựa trên ontology cho CSDLQg về TNMT đã được xây dựng theo phương pháp đề xuất trong chương 2 cho miền tài nguyên và môi trường bao gồm các khái niệm chung của miền tài nguyên và môi trường và tích hợp 02 ontology đã được xây dựng cho hai lĩnh vực đo đạc bản đồ và khí tượng thủy văn. Tổng số các khái niệm của ontology tích hợp này khoảng 111.150 khái niệm.

Cơ sở dữ liệu

Dữ liệu được sử dụng cho hệ thống tìm kiếm ngữ nghĩa dựa trên ontology cho CSDLQg về TNMT bao gồm các dữ liệu dạng văn bản, dữ liệu dạng bảng, dữ liệu không gian được lưu trữ trong các hệ thống cơ sở dữ liệu thuộc hệ thống Cơ sở dữ liệu quốc gia về tài nguyên và môi trường. Hệ thống cơ sở dữ liệu quốc

gia về tài nguyên môi trường đã được xây dựng trong dự án Chính phủ về xây dựng CSDLQg về TNMT do Bộ Tài nguyên và Môi trường chủ trì thực hiện.

Kết quả thử nghiệm

Sau quá trình triển khai, cài đặt và chạy thử nghiệm, đồng thời thực hiện việc so sánh với Phần mềm hiện tại đang triển khai cho CSDLQg về TNMT (NRESearch), chúng tôi đưa ra một số đánh giá sau:

- Phần mềm tìm kiếm ngữ nghĩa (VnNRESS) đã được thiết kế, lập trình và triển khai bảo đảm các yêu cầu chức năng, phi chức năng đã được xác định trong Thuyết minh của đề tài.

- So sánh với Phần mềm tìm kiếm hiện đang được triển khai cho hệ thống CSDLQg về TNMT (nredb.ciren.vn), phần mềm tìm kiếm ngữ nghĩa của Đề tài đã cho kết quả tốt hơn, tập trung ở một số vấn đề sau:

+ Gợi ý từ khóa tìm kiếm: đã thực hiện việc gợi ý theo từ điển từ tiếng Việt, trong khi đó đối với phần mềm NRESearch việc gợi ý từ khóa tìm kiếm chưa đúng. Ngoài ra, phần mềm VnNRESS có chức năng gợi ý từ khóa trong trường hợp người dùng gõ tiếng Việt không dấu.

+ Xử lý tiếng Việt không dấu: phần mềm VnNRESS đã thực hiện việc bỏ dấu tiếng Việt cho hầu hết các từ tiếng Việt không dấu mà người dùng nhập vào, trong khi đó phần mềm NRESearch chưa có chức năng này.

+ Vấn đề tách từ và xác định từ khóa: Đối với phần mềm NRESearch đơn giản là việc tách chuỗi tìm kiếm của người dùng nhập vào theo khoảng trắng (theo từ đơn). Trong khi đó, phần mềm VnNRESS thực hiện việc tách theo các từ có nghĩa trong từ điển, có thể tách theo các từ khóa lồng nhau (ví dụ: "thành phố hồ chí minh" có thể hiểu là 3 từ khóa "thành phố hồ chí minh", "thành phố" và "hồ chí minh"). Với việc tách thành các chuỗi từ khóa có nghĩa như trên, kết quả tìm kiếm sẽ phù hợp hơn với yêu cầu cần tìm kiếm của người dùng.

+ Vấn đề loại bỏ từ dừng: Với bộ từ điển từ dừng kế thừa từ các công trình nghiên cứu trước đó, việc loại bỏ các từ dừng (các từ ít có nghĩa) sẽ giúp cho kết quả tìm kiếm phù hợp hơn và lọc được bớt đi các kết quả không có có nghĩa.

+ Vấn đề sắp xếp kết quả tìm kiếm: Qua quá trình thử nghiệm với một loạt

các từ khóa tìm kiếm, phần mềm VnNRESS sắp xếp thứ tự ưu tiên hiển thị các kết quả tìm kiếm phù hợp hơn với phần mềm NRESearch.

+ Vấn đề xử lý các toán tử tìm kiếm: Phần mềm VnNRESS đã hỗ trợ hầu hết các toán tử tìm kiếm cơ bản và cho độ chính xác hơn đối với phần mềm NRESearch.

+ Vấn đề về độ chính xác kết quả tìm kiếm: Qua các thử nghiệm ở trên, phần mềm VnNRESS đã đưa ra danh sách các kết quả phù hợp hơn với yêu cầu tìm kiếm trong nội dung tìm kiếm của người dùng. Qua đó đã lọc, loại đi nhiều các kết quả tìm kiếm không phù hợp đã được hiển thị trong phần mềm NRESearch.

- Ngoài ra, phần mềm VnNRESS đã được bổ sung một số chức năng sau:

+ Hỗ trợ gợi ý nội dung tìm kiếm khác, phù hợp hơn so với nội dung tìm kiếm người dùng đã nhập.

+ Có chức năng nhận dạng thực thể dạng tên trong nội dung tìm kiếm của người dùng qua việc sử dụng thông tin về địa danh và thông tin về đơn vị hành chính: tỉnh, huyện, xã.

+ Hỗ trợ việc chuẩn xác hóa xác định từ khóa thông qua mối liên quan giữ các từ (ví dụ: các tên địa danh thường đi liền với nhau; thông tin tỉnh/huyện/xã thường đi liền với nhau, …).

+ Hỗ trợ việc xác định các thông tin về địa danh, đưa ra câu trả lời chính xác về tên, vị trí địa lý, vị trí trên bản đồ của các địa danh xuất hiện trong câu tìm kiếm của người dùng.

+ Hỗ trợ việc tìm kiếm theo các từ đồng nghĩa, các từ liên quan (các khái niệm mức trên, mức dưới trong ontology).

- Về chất lượng của phần mềm VnNRESS: Phần mềm đã được Trung tâm Kiểm định sản phẩm CNTT của Cục Công nghệ thông tin và dữ liệu tài nguyên môi trường kiểm thử, kiểm tra kỹ bảo đảm các chức năng hoạt động chính xác, ổn định và đạt yêu cầu chất lượng.

Về thời gian tìm kiếm: So sánh với phần mềm NRESearch, tuy mất nhiều thời gian trong việc phân tích câu hỏi của người dùng, xác định danh mục từ

khóa phù hợp (trên cơ sở từ điển hơn 110.000 từ và các thông tin về địa danh, hành chính) nhưng thời gian thực hiện tìm kiếm là chấp nhận được và bảo đảm yêu cầu.

KẾTLUẬN

I. Những kết quả chính của luận án

Luận án tham gia vào dòng nghiên cứu về học ontology trên thế giới và đạt được ba đóng góp chính là đề xuất được 1 quy trình xây dựng ontology cho miền tài nguyên và môi trường và hai mô hình học ontology phục vụ cho nâng cấp, làm giàu ontology.

Thứ nhất, quy trình xây dựng ontology đã đề xuất bao gồm bảy bước và một giải pháp chia làm ba giai đoạn cho việc xây dựng một ontology cho miền tài nguyên và môi trường (trong đó có chứa miền khí hậu Việt Nam). Quy trình và giải pháp này vừa tổng hợp các kết quả nghiên cứu của luận án vừa cung cấp một phương án thực thi các kết quả nghiên cứu vào thực tiễn [NNVu2] [NNVu6].

Thứ hai, Mô hình học ontology sử dụng phương pháp kết hợp giữa học máy Maximum Entropy và Beam Search nhận dạng thực thể miền [NNVu1] đã cho kết quả tốt không những với bài toán nhận dạng thực thể thông thường mà còn đối với nhận dạng thực thể lồng nhau.

Thứ 3, Mô hình học sâu suốt đời mức ký tự kết hợp các phương pháp học sâu và trường điều kiện ngẫu nhiên cùng với đặc trưng tiền tố (đặc trưng học suốt đời) để nhận dạng thực thể qua đó áp dụng trích xuất các khái niệm, quan hệ, thể hiện từ các văn bản miền phục vụ nâng cấp, mở rộng ontology cần xây dựng cho miền [NNVu4][NNVu5].

Ngoài ra, về phương diện ứng dụng, luận án đã đề xuất hai mô hình kiến trúc hệ thống tìm kiếm ngữ nghĩa dựa trên ontology và hệ thống hỏi đáp dựa trên ontology và tiến hành xây dựng các hệ thống thử nghiệm. Hệ thống tìm kiếm ngữ nghĩa dựa trên ontology cho CSDLQg về tài nguyên môi trường và Hệ thống hỏi đáp pháp luật dựa trên ontology ngành tài nguyên và môi trường hiện nay đã được triển khai vận hành, mang lại nhiều hiệu quả về quản lý và nghiệp vụ cho Bộ Tài nguyên và Môi trường [NNVu3].

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 các kết quả nghiên cứu từ luận án có tiềm năng ứng dụng thực tiễn cao.

II. Hạn chế của luận án

Trong quá trình triển khai các mô hình, một số nghiên cứu trong luận án chưa được tiến hành một cách công phu, thấu đáo để rút ra kết luận bổ ích, cụ thể là:

Một là, miền ứng dụng mới áp dụng để xây dựng ontology miền cho miền tài nguyên và môi trường. Miền tài nguyên và môi trường là miền có phạm vi rộng bao gồm 9 lĩnh vực, do đó hệ thống các khái niệm rất phức tạp và tồn tại nhiều khái niệm ở dạng tiếng Anh và hiện còn có nhiều cách hiểu khác nhau, do đó có một số khái niệm, thuật ngữ được hiểu trên cơ sở tham khảo các từ điển thuật ngữ chuyên ngành và một số chuyên gia miền nên có thể có một vài khái niệm chưa phản ánh được chính xác 100% khi chuyển sang tiếng Việt.

Hai là, một trong những sản phẩm của luận án là ontology miền tài nguyên và môi trường, cần phải có thêm thời gian để các chuyên gia trong các lĩnh vực của ngành tài nguyên và môi trường, chỉnh sửa, cập nhật để nâng cao chất lượng và độ tin cậy của ontology miền này.

Ba là, và là điều quan trọng nhất, luận án chưa tiến hành phân tích đủ sâu các kỹ thuật nâng cấp, làm giàu ontology và các kỹ thuật đề xuất trong luận án mới tập trung chủ yếu vào một số kỹ thuật dựa trên xử lý ngôn ngữ tự nhiên và thống kê, các kỹ thuật nâng cấp ontology dựa trên logic chưa được thực hiện.

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 kỹ thuật nâng cấp, làm giàu ontology.

Một là, nghiên cứu, tìm kiếm, chọn lựa thêm các ontology miền tài nguyên và môi trường cho 9 lĩnh vực: đất đai; tài nguyên nước; tài nguyên khoáng sản, địa chất; môi trường; khí tượng thủy văn; biến đổi khí hậu; đo đạc và bản đồ; quản lý tổng hợp tài nguyên và bảo vệ môi trường biển và hải đảo và viễn thám có chất lượng cao trên thế giới để tích hợp, nâng cấp với ontology hiện có.

*Hai là,*nghiên cứu các kỹ thuật nâng cấp ontology dựa trên logic để có các cải tiến, thử nghiệm và áp dụng cho bài toán xây dựng và hoàn thiện ontology cho miền tài nguyên và môi trường.

Ba là, tiếp tục các hướng nghiên cứu sử dụng ontology miền tài nguyên và môi trường nhằm nâng cao hiệu quả của các bài toán như: tìm kiếm thông tin, hỏi đáp tự động, khai phá văn bản, …phục vụ công tác nghiệp vụ và công tác quản lý tại Bộ Tài nguyên và Môi trường.

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

1. [NNVu1] Cam-Van Thi Nguyen, Thai-Son Pham, Thi-Hong Vuong, Ngoc Vu Nguyen, Mai-Vu Tran (2016). *DSKTLAB-NER: Nested Named Entity Recognition in Vietnamese Text*. VLSP 2016.
2. [NNVu2] Nguyễn Ngọc Vũ (2017). *Xây dựng ontology tài nguyên và môi trường phục vụ tích hợp dữ liệu và tìm kiếm ngữ nghĩa*. Tạp chí Tài nguyên và Môi trường, 30-32 (2017).
3. [NNVu3] Ngoc-Vu Nguyen, Thi-Lan Nguyen, Cam-Van Nguyen Thi, Mai-Vu Tran, Quang-Thuy Ha (2019). *A Character-Level Deep Lifelong Learning Model for Named Entity Recognition in Vietnamese Text*. ACIIDS (1) 2019: 90-102. (***Scopus, DBLP).***
4. [NNVu4] Ngoc-Vu Nguyen, Hong-Son Bui, Quang-Thuy Ha (2019). *Ontology-Based Semantic Search for National Database of Natural Resources and Environment*. INISCOM 2019: 155-164. (***Scopus, DBLP).***
5. [NNVu5] Ngoc-Vu Nguyen, Thi-Lan Nguyen, Cam-Van Nguyen Thi, Mai-Vu Tran, Tri-Thanh Nguyen, Quang-Thuy Ha (2019). *Improving Named Entity Recognition in Vietnamese Texts by a Character-Level Deep Lifelong Learning Model*. Vietnam J. Computer Science 6(4): 471-487 (2019).. (***Scopus, DBLP).***
6. [NNVu6] Ngoc-Vu Nguyen, Hai-Chau Nguyen, Quang-Thuy Ha. Developing a Domain Ontology for Natural Resources and Environment. The 6th NAFOSTED Conference on Information and Computer Science (NICS), in press. ***(Scopus, DBLP)***.

TÀI LIỆU THAM KHẢO

[1] R. Arp, B. Smith and A. D. Spear, Building Ontologies with Basic Formal Ontology, The MIT Press, 2015.

[2] J. Cullen and A. Bryman, "The knowledge acquisition bottleneck: time for reassessment?," *Expert Systems,* vol. Vol 5 No 3, pp. 216-225, 1988.

[3] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood and H. M. Abbasi, "A survey of ontology learning techniques and applications," *Database,* vol. 2018, p. bay101, 2018.

[4] Natalya F. Noy, Ontology Development 101: A Guide to Creating Your First Ontology, Stanford University, 2001.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[6] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing,* vol. 45, no. 11, pp. 2673-2681, 1997.

[7] T. Jo, "Taxonomy Generation," in *Text Mining*, 2018, pp. 319-340.