



DOI:10.22144/ctu.jsi.2017.018

XÂY DỰNG ONTOLOGY TỰ ĐỘNG TỪ BẢNG CHÚ GIẢI

Trần Công Ân¹, Tống Thị Ngọc Mai¹ và Lê Thị Thu Lan²

¹Khoa Công nghệ Thông tin & Truyền thông, Trường ĐHTC

²Khoa Kỹ Thuật Công nghệ, Trường Đại học Tây Đô

Thông tin chung:

Ngày nhận bài: 15/09/2017

Ngày nhận bài sửa: 10/10/2017

Ngày duyệt đăng: 20/10/2017

Title:

Learning lightweight ontology from glossary

Từ khóa:

Bảng chú giải, biểu thức chính quy, ontology, tự động, WordNet

Keywords:

Glossary, learning, ontology, WordNet, regular expression

ABSTRACT

Ontology is an advanced knowledge representation formalism. It allows reusing and sharing vocabularies between applications and plays an important role in Semantic Web. However, ontology development is complicated and time-consuming. Therefore, in this paper, an approach to constructing lightweight ontology from glossary and the WordNet was proposed. This approach based on linguistics techniques such as regular expression and Link Grammar. The experiment on the Internet Movie Database glossary showed a promising result that the proposed approach produced an ontology with more than 600 concepts and 200 relationships. However, the results still existed some limitations that required further improvements.

TÓM TẮT

Ontology là một hình thức biểu diễn tri thức cho phép chia sẻ giữa các ứng dụng và đóng vai trò rất quan trọng đối với web ngữ nghĩa. Việc xây dựng ontology thủ công tương đối phức tạp và mất thời gian. Do đó, trong nghiên cứu này, chúng tôi đề xuất một phương pháp xây dựng một ontology gọn nhẹ (light-weighted ontology) dựa trên bảng chú giải (glossary) kết hợp với cơ sở dữ liệu từ vựng WordNet và một số kỹ thuật trong xử lý ngôn ngữ tự nhiên như biểu thức chính quy, Link Grammar. Phương pháp này được thực nghiệm trên tập dữ liệu IMDB và đã xây dựng được một ontology với hơn 600 khái niệm và 200 quan hệ giữa các khái niệm. Kết quả cho thấy phương pháp được đề xuất là khả thi, cho phép xác định các khái niệm và một số quan hệ giữa chúng. Tuy nhiên, phương pháp vẫn còn một số hạn chế như phát hiện thiếu một số quan hệ giữa các khái niệm, đòi hỏi phải có thêm một số cải tiến khác để đạt được độ chính xác cao hơn.

Trích dẫn: Trần Công Ân, Tống Thị Ngọc Mai và Lê Thị Thu Lan, 2017. Xây dựng ontology tự động từ bảng chú giải. Tạp chí Khoa học Trường Đại học Cần Thơ. Số chuyên đề: Công nghệ thông tin: 133-139.

1 GIỚI THIỆU

Ontology là một trong các hình thức biểu diễn tri thức tiên tiến nhất hiện nay. Với hình thức biểu diễn tri thức này, mô hình các khái niệm và quan hệ giữa các khái niệm trong miền tri thức cho phép các tri thức có thể được sử dụng lại cũng như được chia sẻ giữa các ứng dụng. Ontology được ứng

dụng rộng rãi trong nhiều lĩnh vực như trí tuệ nhân tạo, truy hồi thông tin... Tuy nhiên, ứng dụng rộng rãi nhất của ontology là trong lĩnh vực web ngữ nghĩa (Semantic Web). Đây chính là nền tảng cung cấp ngữ nghĩa cho dữ liệu, cho phép dữ liệu có thể được hiểu bởi máy tính.

Do đây là một công nghệ nền tảng của web ngữ nghĩa, nhu cầu xây dựng các ontology là rất lớn.

Tuy nhiên, việc xây dựng các ontology cho một miền tri thức một cách thủ công mất rất nhiều thời gian, đòi hỏi nhiều nhân lực và cần sự hỗ trợ từ các chuyên gia về lĩnh vực đó. Có nhiều nghiên cứu đề xuất các phương pháp để tăng tốc độ và hiệu quả của việc xây dựng ontology (sẽ được giới thiệu trong phần tiếp theo của bài báo). Ý tưởng cơ bản của các phương pháp này là tự động hoặc bán tự động hoá việc xây dựng các ontology từ các nguồn dữ liệu trong cùng miền tri thức.

Trong bài báo này, chúng tôi sẽ đề xuất một phương pháp để xây dựng một ontology gọn nhẹ (lighweight ontology) dựa trên các bảng chú giải (glossary) của miền tri thức tương ứng. Bảng chú giải của một miền tri thức chứa các thuật ngữ, khái niệm (concept) và định nghĩa (definition) cho các thuật ngữ, khái niệm đó. Đây là một nguồn dữ liệu có cấu trúc, vì vậy việc sử dụng các bảng chú giải sẽ dễ dàng hơn so với các nguồn dữ liệu không có cấu trúc. Ngoài ra, các bản chú giải hiện khá phong phú, có sẵn cả dạng ngoại tuyến (offline, ví dụ như sách) lẫn trực tuyến (online, ví dụ như trên các trang web). Do đó, các bảng chú giải có thể được sử dụng như một nguồn dữ liệu chính để xây dựng các ontology. Ngoài ra, để làm phong phú hơn, tăng độ bao phủ của ontology, trong nghiên cứu này, chúng tôi cũng đề xuất sử dụng thêm nguồn cơ sở dữ liệu từ vựng WordNet.

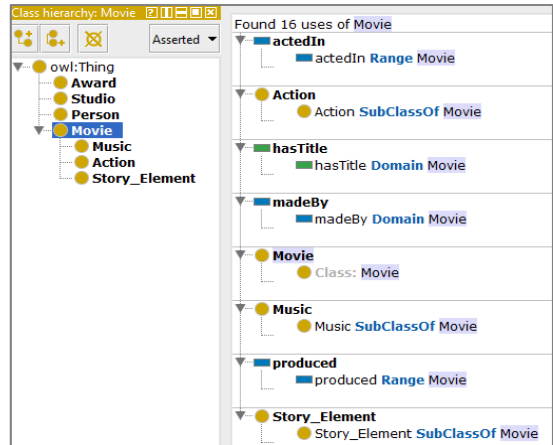
Bài báo này được tổ chức như sau: Phần 2 giới thiệu về ontology và các phương pháp xây dựng ontology đã được đề xuất; ở phần 3, chúng tôi mô tả phương pháp xây dựng ontology tự động dựa trên bảng chú giải do chúng tôi đề xuất và thuật toán cụ thể cho phương pháp này; trong phần 4, chúng tôi sẽ trình bày kết quả thực nghiệm trên tập dữ liệu của IMDB; cuối cùng, thảo luận về kết quả đạt được và hướng phát triển sẽ được trình bày trong phần 5.

2 ONTOLOGY VÀ CÁC PHƯƠNG PHÁP XÂY DỰNG ONTOLOGY

2.1 Ontology và các phương pháp thể hiện tri thức

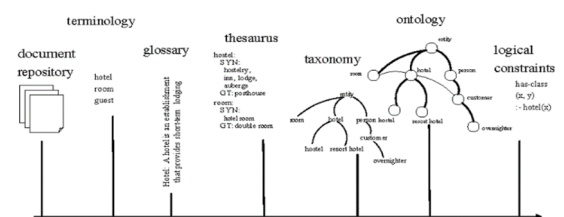
Ontology là một đặc tả chính qui (formal) và tường minh (explicit) của các khái niệm được chia sẻ (T. Gruber, 1993). Một ontology có thể được trực quan hóa bằng một đồ thị có hướng với các đỉnh là các khái niệm và các cạnh biểu diễn mối quan hệ giữa các khái niệm. Đây là một trong các hình thức biểu diễn tri thức chính qui rộng rãi nhất, là nền tảng của web ngữ nghĩa (S. Bechhofer, 2009, T. Berners-Lee, 2001). Hình thức biểu diễn tri thức này độc lập với ngôn ngữ tự nhiên và không sử dụng các tri thức liên quan đến từ vựng (lexical) của ngôn ngữ tự nhiên. 0 minh họa một

phần của ontology về phim ảnh trong đó có các khái niệm như *Award*, *Movie*, *Person*,... và các mối quan hệ giữa các khái niệm như *actedIn*, *madeBy*, *producedBy*, *hasTitle*,...



Hình 1: Một phần của ontology về phim ảnh được biểu diễn trực quan hóa trong Protégé

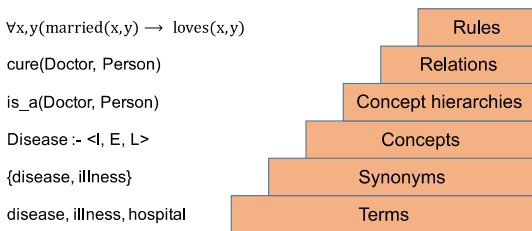
Vị trí của ontology trong các cấp độ biểu diễn tri thức được mô tả trong 0. Đối với hình thức biểu diễn tri thức bằng các tài liệu (document repository), tri thức được biểu diễn bằng ngôn ngữ tự nhiên và không có yêu cầu hay ràng buộc về cấu trúc (không cấu trúc). Đây có thể được xem là cấp độ biểu diễn tri thức thấp nhất và hình thức biểu diễn tri thức này chỉ có thể được hiểu bởi con người còn máy tính thì không thể “hiểu” và xử lý trực tiếp. Ngược lại, cấp độ biểu diễn tri thức cao nhất là các ontology đầy đủ (heavy-weighted ontology). Cấp độ biểu diễn tri thức này sử dụng các luật logic (logical rule), hay còn gọi là các tiên đề (axioms) để biểu diễn tri thức. Điều này làm cho tri thức được biểu diễn trở nên có cấu trúc và do đó có thể được “hiểu” và xử lý trực tiếp bởi máy tính. Ngoài ra, việc ứng dụng logic trong biểu diễn tri thức còn cho phép thực hiện các suy luận (reasoning) trên tập tri thức này. Các cấp độ biểu diễn tri thức khác như thuật ngữ (terminology), bảng chú giải (glossary) và từ điển (thesaurus) được gọi là các hình thức từ vựng được “kiểm soát” (controlled vocabulary). Ví dụ minh họa cho từng cấp độ biểu diễn tri thức có thể được tham khảo trong 0.



Hình 2: Các cấp độ biểu diễn tri thức (G. Miller et al., 1990)

2.2 Các phương pháp xây dựng ontology

Có hai phương pháp xây dựng ontology là xây dựng thủ công và “học” (tự động, bán tự động). Xây dựng ontology thủ công thường mất nhiều thời gian và đòi hỏi phải có các chuyên gia trong lĩnh vực đó. Do đó, phương pháp này chỉ phù hợp để xây dựng các ontology cho các miền tri thức nhỏ, giới hạn và ít thay đổi. Đối với các ontology lớn hay thường thay đổi thì cần có phương pháp khả thi hơn, tiết kiệm thời gian, nhân lực và đáp ứng với các thay đổi tốt hơn, đó chính là phương pháp xây dựng một cách tự động hay bán tự động hay còn gọi là “học” ontology (ontology learning). Học ontology là quá trình xác định các thuật ngữ (term), các khái niệm (concept), các quan hệ phân loại hay cấp bậc (taxonomy relation) và quan hệ không cấp bậc (non-taxonomy relation), các tiên đề (axiom). Đây chính là các thành phần của một ontology. Tuy nhiên, tùy vào từng cấp độ chi tiết, vào qui mô của ontology mà quá trình học có thể chỉ xác định một số trong các thành phần trên. Vị trí của các thành phần trên trong một ontology được gọi là “Ontology Learning Layer Cake” và được mô tả trong 0.



Hình 3: Ontology Learning Layer Cake (S. Bird et al., 2008)

Việc học các thành phần trong Ontology Learning Layer Cake càng lên cao càng phức tạp. Ở mức độ thấp nhất là các thuật ngữ, là thành phần cơ bản nhất của ontology. Một thuật ngữ có thể là một từ đơn, từ kép,... mô tả cho tri thức trong một lĩnh vực cụ thể. Ví dụ, một số thuật ngữ trong lĩnh vực y tế như “bệnh viện” (hospital), “bệnh” (disease, illness), “thuốc” (medicine),... Đồng nghĩa là một nhóm các thuật ngữ có cùng nghĩa với nhau, ví dụ {disease, illness}. Khái niệm là các thuật ngữ có gán nhãn và các khái niệm đồng nghĩa của nó. Mức kế tiếp của việc xây dựng ontology là xác định các quan hệ giữa các khái niệm. Quan hệ cấp bậc là quan hệ “là” (is-a), được dùng để xây dựng cây phân cấp khái niệm (hierarchy). Ví dụ, một “bác sĩ” (Doctor) là một “con người” (Person). Còn loại quan hệ không cấp bậc là các quan hệ giữa các khái niệm ngoài quan hệ cấp bậc. Ví dụ, quan hệ “chữa bệnh” (cure) là quan hệ không cấp bậc giữa “bác sĩ” và “bệnh nhân”. Quan hệ cấp bậc thường để xác định hơn quan hệ không cấp bậc vì

nó có thể được nhận dạng dễ dàng hơn. Ngược lại, quan hệ không cấp bậc thì thường khó xác định hơn vì mỗi quan hệ này thường là không tường minh và đa dạng hơn. Ở mức độ cao nhất của Ontology Learning Layer Cake là các luật logic hay tiên đề. Các luật logic này được định nghĩa trên các khái niệm và quan hệ giữa các khái niệm. Chúng được dùng mô tả những ràng buộc phức tạp trên các khái niệm hoặc các quan hệ. Các ràng buộc này cho phép kiểm tra tính đúng đắn của ontology cũng như giảm kích thước (số lượng các thành phần) của ontology vì một số thành phần không cần khai báo tường minh trong ontology mà có thể được suy luận từ các luật này. Một ví dụ về luật logic trong ontology là: $\forall x,y(\text{cưới}(x,y) \rightarrow \text{yêu}(x,y))$ (với hai người bất kỳ x, y nếu x cưới y thì có nghĩa là x yêu y). Như vậy, giả sử trong ontology đã chứa tri thức cưới(x,y) thì ta có thể suy luận ra là x và y yêu nhau mà không cần phải thêm tri thức này vào ontology.

Hầu hết các nghiên cứu hiện tại sử dụng tập ngữ liệu (text corpus) của miền tri thức kết hợp với các kỹ thuật như máy học hoặc các kỹ thuật trong xử lý ngôn ngữ tự nhiên,... để xây dựng hay học ontology. Các phương pháp này có thể được phân làm 3 loại: phương pháp dựa trên thống kê (statistic-based), phương pháp dựa trên logic và phương pháp dựa trên xử lý ngôn ngữ tự nhiên. Một số nghiên cứu cụ thể kết hợp nhiều phương pháp lại với nhau. Wong et al. (2012) đã thực hiện một nghiên cứu tổng quan về các phương pháp xây dựng ontology tự động.

Trong các phương pháp xây dựng ontology tự động trên, phương pháp dựa trên thống kê và ngôn ngữ tự nhiên được sử dụng rộng rãi hơn phương pháp dựa trên logic. Phương pháp dựa trên thống kê dựa vào 1 tiên đề là các từ xuất hiện cùng nhau của các từ vựng thường có nghĩa là chúng có mối liên hệ với nhau. Phân cụm (clustering) là 1 kỹ thuật phổ biến dùng để phân các thuật ngữ vào các nhóm dựa vào độ đo tương đồng (similarity measure) (K. Linden and J. Piitulainen, 2004). Trong nghiên cứu của H. Fotzo và P. Gallinari (2004), các tác giả đề xuất một phương pháp để xây dựng mối quan hệ phân cấp bằng cách sử dụng xác suất có điều kiện của sự xuất hiện thuật ngữ trong tài liệu. Cho hai thuật ngữ x và y, nếu $P(x|y) < P(y|x)$ và $P(x|y) > t$ với t là một ngưỡng cho trước, $P(x|y)$ là xác suất xuất hiện thuật ngữ x khi có thuật ngữ y thì x và y có quan hệ với nhau. Một phương pháp dựa trên thống kê nữa là sử dụng độ đo TF-IDF để đo tần suất xuất hiện của thuật ngữ trong các tập ngữ liệu với qui mô khác nhau (các tập ngữ liệu chung, các tập ngữ liệu trong một miền tri thức cụ thể,...) (G. Salton and

C. Buckley, 1988). Các thuật ngữ có quan hệ với nhau thường xuất hiện cùng nhau trên nhiều tập ngữ liệu khác nhau. Hạn chế của phương pháp dựa trên thống kê là không ứng dụng được ngữ nghĩa cũng như các đặc điểm của ngôn ngữ trong quá trình xây dựng ontology.

Phương pháp xây dựng ontology tự động dựa trên các kỹ thuật xử lý ngôn ngữ tự nhiên được áp dụng rộng rãi hơn cả vì nó có thể khắc phục những hạn chế của phương pháp dựa trên thống kê. Ngoài ra, phương pháp này còn vận dụng được các công cụ xử lý ngôn ngữ tự nhiên rất mạnh đã được phát triển. Ví dụ như TreeTagger (G. Salton and C. Buckley, 1988) và Link Grammar Parser (D. Temperley and D. Sleator, 1993) là các công cụ gắn nhãn từ loại (POS tagging) và phân tích ngữ pháp rất mạnh. Hoặc NLTK (Natural Language Toolkit) là một bộ công cụ toàn diện cho các tác vụ xử lý ngôn ngữ tự nhiên. Phân tích cú pháp câu có thể giúp xác định các thuật ngữ cũng như các quan hệ giữa các khái niệm. Ví dụ, kết quả một phân tích cấu trúc “thuật ngữ” – “động từ” – “thuật ngữ” thì “động từ” có thể được coi là một ứng cử viên cho một quan hệ. Các cơ sở dữ liệu từ vựng (lexical database) như WordNet (G. Miller *et al.*, 1990) cũng rất hữu ích để tìm kiếm các khái niệm và quan hệ đã được định nghĩa trước (synonym, hyponym, hypernym, meronym,...) (W. Zhou *et al.*, 2006).

Mặc dù tập ngữ liệu (text corpus) vẫn được sử dụng như là nguồn tài nguyên chính bởi đa số các phương pháp xây dựng ontology tự động, có nhiều nghiên cứu đang đề xuất việc sử dụng các nguồn dữ liệu có cấu trúc để xây dựng hoặc “làm giàu” ontology. Ví dụ, Wikipedia, một trong những cơ sở tri thức trực tuyến lớn nhất, là một nguồn tài nguyên có giá trị để trích xuất quan hệ giữa các thuật ngữ. Liu *et al.* (2008) đã sử dụng hệ thống phân loại của Wikipedia và hộp thông tin trên hệ thống này để trích xuất các bộ ba (triple, ví dụ <Can Tho, city, Viet Nam >), dùng để xây dựng ontology. Ngoài ra, cũng có một số nghiên cứu sử dụng bảng chú giải để xây dựng ontology (J. Hilera *et al.*, 2010; M. Li *et al.*, 2005) hoặc “làm giàu” ontology đã có sẵn (R. Navigli and P. Verladi, 2008). Tuy nhiên, trong các nghiên cứu này thì chỉ sử dụng duy nhất nguồn dữ liệu là ontology nên ontology tạo được còn hạn chế.

3 XÂY DỰNG ONTOLOGY DỰA TRÊN BẢNG CHÚ GIẢI

Trong phần này, chúng tôi sẽ đề xuất một phương pháp xây dựng ontology gọn nhẹ dựa trên bảng chú giải. Bảng chú giải là một danh sách các khái niệm và định nghĩa của khái niệm đó. Ví dụ,

dưới đây là một số mục trong bảng chú giải của IMDB (The Internet Movie Database):

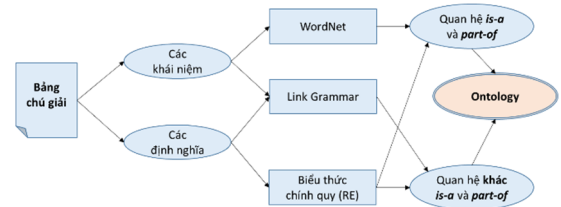
Agent

A person responsible for the professional business dealings of an actor, director, or other artist. An agent typically negotiates the contracts on behalf of the actor or director, and often has some part in selecting or recommending roles for their client.

Art Director

The person who oversees the artists and craftspeople who build the sets. See also production designer, set designer, set director, leadman, and swing gang.

Trong ví dụ trên, “Agent” và “Art director” là các khái niệm và “A person responsible for...” và “The person who oversee...” là định nghĩa tương ứng cho các khái niệm này. Do đó, mỗi khái niệm trong bảng chú giải chính là một khái niệm trong ontology. Việc xác định các khái niệm này khá dễ dàng. Như vậy, trọng tâm của phương pháp xây dựng ontology tự động là xác định mối quan hệ giữa các khái niệm. Ngoài ra, xác định các khái niệm mở rộng trong bảng chú giải cũng đòi hỏi sử dụng các kỹ thuật trong xử lý ngôn ngữ tự nhiên hoặc máy học. Trong nghiên cứu này, chúng tôi đề xuất sử dụng các biểu thức chính quy (regular expression) kết hợp với các kỹ thuật phân tích ngữ pháp POS tagger trong xử lý ngôn ngữ tự nhiên để tìm ra các khái niệm mở rộng cũng như các quan hệ giữa các khái niệm. Mô hình xây dựng ontology tự động được trình bày trong 0.



Hình 4: Mô hình xây dựng ontology tự động từ bảng chú giải

3.1 Biểu thức chính quy

Bảng chú giải thường chứa các tham chiếu chéo (cross-references) giữa các khái niệm. Do đó, các biểu thức chính quy sẽ rất hữu ích trong việc tự động nhận dạng và phân loại các quan hệ. Một số tham chiếu chéo như:

– “contrast with” hoặc “disjoint with”: thể hiện quan hệ *trái nghĩa* giữa hai khái niệm. Ví dụ:

arcade: A series of arches supported by columns or piers. **Contrast with** *colonnade*

– “synonym”, “as known as”, “to be known as” hoặc “see”: thể hiện quan hệ *đồng nghĩa* giữa các khái niệm.

barrel vault, also known as a tunnel vaultor: an architectural element formed by the extrusion of a single curve along a given distance

– “see also”: thể hiện các khái niệm *có liên quan* với nhau.

anamorphic: An optical system which has different magnifications in the vertical and horizontal dimensions of the picture. **See also aspect ratio, contrast with spherical.**

– “sub class of”: thể hiện mối quan hệ *đặc trưng hóa*. (quan hệ *cha-con* giữa các khái niệm).

eutherians: a **subclass of mammals** which give birth to well-developed young. Humans are part of this subclass.

Ngoài ra, một số mẫu cú pháp ngôn ngữ khác cũng rất hữu ích trong việc xác định các quan hệ giữa các khái niệm. Ví dụ, “*concept₁ such as concepts₂*” ngụ ý rằng *concept₂* là 1 khái niệm con của *concept₁*. Do đó, bằng cách sử dụng các quan hệ đơn giản này giữa các khái niệm, cho phép xác định nhiều quan hệ trong ontology.

3.2 WordNet

WordNet là một cơ sở dữ liệu từ vựng lớn nhất của tiếng Anh, được sử dụng như là một nguồn tài nguyên quan trọng trong rất nhiều ứng dụng về xử lý ngôn ngữ tự nhiên và trong các lĩnh vực khác có liên quan. WordNet cho phép ta truy xuất một cách dễ dàng đến các khái niệm và một tập quan hệ rất phong phú giữa các khái niệm như synonyms,

hyponymy, hypernymy, meronymy,... Do đó, cho một cặp khái niệm, ta có thể kiểm tra xem một mối quan hệ nào đó có tồn tại giữa chúng hay không. Nói cách khác, cho một khái niệm *C*, nếu một khái niệm *C'* nằm trong tập *hypernym* hoặc *meronym* của *C*, hoặc các quan hệ ngữ nghĩa khác thì ta có thể thêm mối quan hệ tương ứng vào ontology.

3.3 Link Grammar

Quan sát cú pháp của một câu, ta có thể dễ dàng thấy được nếu hai khái niệm xuất hiện trong cùng một câu thì động từ liên kết hai khái niệm này thường thể hiện cho quan hệ giữa hai khái niệm. Để thực hiện phân tích câu và tìm động từ thể hiện mối quan hệ giữa các khái niệm, chúng ta cần phải phân tích cú pháp (syntactic) và phụ thuộc (dependency). Trong nghiên cứu này, chúng tôi đề xuất sử dụng Link Grammar (D. Temperley and D. Sleator, 1993), một trong những kỹ thuật sử dụng rộng rãi nhất cho việc phân tích cú pháp câu. Link Grammar không chỉ tạo ra cây cú pháp như POS tagger mà còn cung cấp thông tin về sự phụ thuộc giữa các cặp từ trong câu dưới dạng liên kết (quan hệ). Trong nghiên cứu này, chúng tôi sử dụng PTQL (L. Tari *et al.*, 2010) thay sử dụng trực tiếp Link Grammar. Ngoài việc hỗ trợ xây dựng cây cú pháp và các liên kết, PTQL còn hỗ trợ lưu trữ các thông tin này vào cơ sở dữ liệu quan hệ và ngôn ngữ truy vấn để chúng ta có thể thao tác (sửa đổi, bổ sung, truy vấn) dữ liệu một cách dễ dàng.

3.4 Giải thuật xây dựng ontology tự động từ bảng chú giải

Giải thuật xây dựng ontology tự động từ bảng chú giải được mô tả bằng mã giả (pseudo code) trong Giải thuật 1.

Giải thuật 1: Xây dựng ontology tự động từ bảng chú giải

Đầu vào: các bảng chú giải trong miền tri thức

Đầu ra: ontology cho miền tri thức tương ứng

BEGIN

1. **ontology** = \emptyset
2. **concepts** = {tập các khái niệm trong các bảng chú giải}
3. **sentences** = {tập các câu trong định nghĩa của các khái niệm trong bảng chú giải}
4. **tax_RE_relations** = RE_concepts(concepts, sentences, taxonomy-RE)
//tìm các quan hệ cấp bậc (taxonomy relations) giữa các khái niệm dựa trên biểu thức chính quy đã định nghĩa sẵn
5. **nonTax_RE_relations** = RE_concepts(concepts, sentences, non-taxonomy-RE)
//tìm các quan hệ không cấp bậc (non-taxonomy relations) giữa các khái niệm dựa trên biểu thức chính quy
6. **hyp_WN_concepts** = WordNet_Hyponym_Query(concepts)
//tìm các cặp khái niệm có quan hệ hypernym với nhau dựa trên WordNet, hyp_WN_concepts = {(c_i, c_j)}
7. **mer_WN_concepts** = WordNet_Meronym_Query(concepts)
//tìm các cặp khái niệm có quan hệ meronym với nhau dựa trên WordNet, mer_WN_concepts = {(c_i, c_j)}
8. **LG_parse_tree_rel** = TPQL(sentences)
//xây dựng cây phân tích cú pháp cho tập sentences, sử dụng TPQL
9. **ontology** = ontology \cup concepts
10. $\forall r(c_1, c_2) \in \{\text{tax_RE_relations} \cup \text{nonTax_RE_relations} \cup \text{LG_parse_tree_rel}\}$:
 ontology = ontology $\cup r$ *//thêm quan hệ r vào ontology (nếu r chưa tồn tại trong ontology)*
 ontology = ontology $\cup r(c_1, c_2)$ *//thêm một thể hiện (instance) của quan hệ r vào ontology*
11. $\forall (c_1, c_2) \in \text{mer_WN_concepts}$: **ontology** = ontology $\cup \text{hyponym}(c_1, c_2)$
 //thêm quan hệ hyponym cho các cặp khái niệm trong tập hyp_WN_concepts
12. $\forall (c_1, c_2) \in \text{mer_WN_concepts}$: **ontology** = ontology $\cup \text{meronym}(c_1, c_2)$
 //thêm quan hệ meronym cho các cặp khái niệm trong tập mer_WN_concepts

END.

4 KẾT QUẢ THỰC NGHIỆM

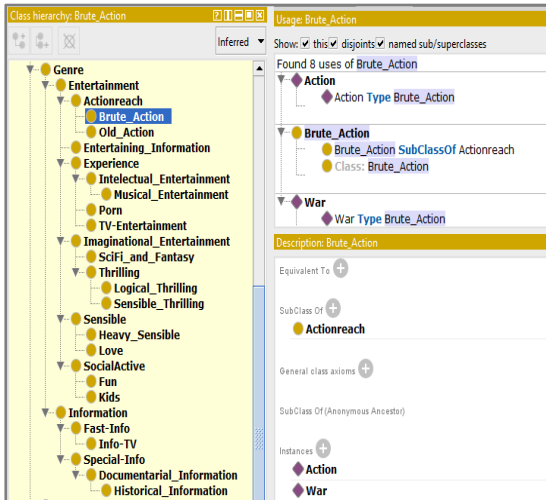
Để minh họa và kiểm chứng phương pháp được đề xuất trong nghiên cứu này, chúng tôi sử dụng một bảng chú giải về phim ảnh IMDB được lấy trực tiếp từ internet tại địa chỉ <http://www.imdb.com/glossary/>. Bảng chú giải này bao gồm 605 thuật ngữ liên quan đến phim ảnh. Do bảng chú giải này có định dạng HTML nên cần phải có bước tiền xử lý để loại bỏ các định dạng như các thẻ HTML, chỉ giữ lại nội dung của bảng chú giải. Ngoài ra, bước tiền xử lý còn phải trích xuất ra các thuật ngữ và các định nghĩa của thuật ngữ để làm đầu vào cho hệ thống.

Kết quả đạt được là hệ thống đề xuất đã tìm được hơn 200 quan hệ giữa các khái niệm. Một phần của ontology này được trình bày trong 0. Tuy nhiên, hầu hết các quan hệ tìm được thuộc dạng quan hệ cấp bậc (taxonomy). Ngoài ra, hầu hết các quan hệ tìm được bởi các biểu thức chính quy (120 quan hệ) và WordNet (60 quan hệ). Có một số lý do giải thích cho việc thiếu các quan hệ không cấp

bậc (non-taxonomy) trong kết quả là nguồn dữ liệu chú giải IMDB tương đối đơn giản, đa số các định nghĩa chỉ từ 1 đến 2 câu. Vì vậy thông tin được cung cấp bởi dữ liệu chú giải là không đủ phong phú để có các quan hệ khác quan hệ cấp bậc. Ngoài ra, trong dữ liệu chú giải này có rất nhiều quan hệ dạng tham chiếu chéo (cross-reference) trong khi thiếu các định nghĩa hoặc định nghĩa rất ngắn gọn. Một trong những lý do nữa là giải thuật hiện tại chưa xử lý được các cụm tính từ có ý nghĩa như một quan hệ như “responsible for”... Tuy nhiên, việc xử lý trường hợp này là khả thi đối với mô hình được đề xuất.

5 KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã đề xuất một mô hình xây dựng ontology gọn nhẹ tự động từ bảng chú giải dựa trên các kỹ thuật xử lý ngôn ngữ tự nhiên. Bên cạnh bảng chú giải, chúng tôi có sử dụng thêm cơ sở dữ liệu từ vựng WordNet để làm giàu thêm các mối quan hệ giữa các khái niệm trong ontology.



Hình 5: Một phần của ontology được xây dựng dựa trên nguồn dữ liệu chú giải IMDB

Ưu điểm của phương pháp này là đơn giản, các kỹ thuật được đề xuất sử dụng đều có tốc độ xử lý nhanh. Ngoài ra, nguồn dữ liệu để xây dựng ontology không đòi hỏi phải quá lớn như trong phương pháp dựa trên thống kê. Tuy nhiên, trong một số trường hợp thì kích thước của ontology thu được còn nhỏ so với phương pháp dựa trên thống kê.

Do đó, hướng phát triển của nghiên cứu này là bổ sung thêm nguồn dữ liệu hỗ trợ như Wikipedia hoặc bổ sung thêm các dạng của biểu thức chính quy để có thể nhận dạng được nhiều loại quan hệ hơn như trong trường hợp tính từ dùng như động từ đã thảo luận bên trên. Ngoài ra, có thể kết hợp thêm các kỹ thuật của logic để trích xuất các tiên đề (axioms) để tiến tới việc xây dựng các ontology đầy đủ thay vì ontology gọn nhẹ như hiện tại.

TÀI LIỆU THAM KHẢO

A. Oliveira, C. Pereira, and A. Cardoso, 2001. Automatic reading and learning from text. In Proceedings of the International Symposium on Artificial Intelligence (ISAI), Kolhapur, India.

D. Temperley and D. Sleator, 1993. Parsing english with a link grammar. In Proceedings of the 3rd International Workshop on Parsing Technologies.

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, 1990. Wordnet: An on-line lexical database. International Journal of Lexicography, 3: 235-244.

G. Salton and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. In Information Processing and Management, volume 24, pp. 513-523.

H. Fotzo and P. Gallinari, 2004. Learning generalization/specialization relations between

concepts application for automatically building thematic document hierarchies.

J. Hilera, C. Pages, J. Martinez, J. Gutierrez, and L. de Marcos, 2010. An evolutive process to convert glossaries into ontologies. Information Technology and Libraries (ITAL), 29:195-204.

J. Tang, H. Leung, Q. Luo, D. Chen, and J. Gong, 2009. Towards ontology learning from folksonomies. In Proceedings of the 21st international Joint conference on Artificial intelligence, IJCAI'09, pages 2089-2094.

K. Linden and J. Piitulainen, 2004. Discovering synonyms and other related words. In Proceedings of the CompuTerm, Geneva, Switzerland.

L. Tari, P. Tu, J. Hakenberg, Y. Chen, T. Son, G. Gonzaler, and C. Baral, 2010. Incremental information extraction using relational databases. IEEE Transactions on knowledge & Data Engineering, 24: 86-99.

M. Li, X. Du, and S. Wang, 2005. Learning ontology from relational database. In Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, volume 6, pages 3410-3415.

Q. Liu, K. Xu, L. Zhang, H. Wang, Y. Yu, and Y. Pan, 2008. Catriple: Extracting triples from wikipedia categories. In Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web, ASWC '08, Springer-Verlag, pages 330-344.

R. Navigli and P. Verladi, 2008. From glossaries to ontologies: Extracting semantic structure from textual definitions.

S. Bechhofer, 2009. OWL: Web ontology language. In Encyclopedia of Database Systems. Springer US, pages 2008-2009.

S. Bird, E. Klein, E. Loper, and J. Baldrige, 2008. Multidisciplinary instruction with the natural language toolkit. In Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics, pages 62-70.

T. Berners-Lee, 2001. The Semantic Web.

T. Gruber, 1993. A translation approach to portable ontology specifications. Knowledge Acquisition, 5:199-220.

The Internet Movie Database (IMDB). <http://www.imdb.com/glossary/>, last access: 6/2017.

W. Wong, W. Liu, and M. Bennamoun, 2012. Ontology learning from text: A look back and into the future. ACM Computing Surveys, 44(4).

W. Zhou, Z. Liu, Y. Zhao, L. Xu, G. Chen, Q.Wu, M. Huang, and Y. Qiang, 2006. A semi-automatic ontology learning based on wordnet and event-based natural language processing. International Conference on Information and Automation, pages 240-244.