

UNIVERSITY OF INFORMATION AND TECHNOLOGY

Khoá luận tốt nghiệp 1-2015

hoàn thành bởi

Chương Đặng - 10520010

Duy Nguyễn - 10520011

với sự hướng dẫn của

TS. Nguyễn Anh Tuấn

Đề tài khóa luận

tốt nghiệp - sinh viên khóa MMTT2010

thuộc

Chuyên ngành phát triển ứng dụng Web và di động

Khoa mạng máy tính và truyền thông

Ngày 4 tháng 1 năm 2015

Lời mở đầu

Thế giới của chúng ta đang liên tục vận động theo chiều hướng tích cực. Đây là nguyên nhân chủ yếu cho sự phát triển và thay đổi hàng ngày của mọi lĩnh vực đời sống, đặc biệt là khoa học công nghệ nói chung, và ngành công nghệ thông tin nói riêng. Hiện nay, hầu hết mọi nơi trên thế giới đều đã biết đến sự có mặt của công nghệ thông tin, máy tính, và kể cả internet. Việc internet ra đời là một sự kiện đã làm thay đổi cả thế giới. Thay thế cho việc gọi điện thoại hàng ngày, chúng ta có thể liên lạc qua internet, với việc có thể thấy được hình ảnh của người đối diện chứ không chỉ riêng giọng nói. Thay thế cho những tờ báo bằng giấy, chúng ta đã có những trang web, với những thông tin đầy đủ hơn, hình ảnh sinh động hơn, và cả những đoạn video minh họa. Những thông tin trên internet được lan truyền với tốc độ chóng mặt, dẫn đến việc những tin tức nóng nhất được cập nhật liên tục trong từng phút từng giây. Cách tiếp cận thông tin của con người thay đổi, nên sự chuyển động của thông tin ngày càng nhanh hơn, và đến mức nào đó, thông tin sẽ không mang theo đủ những gì mà con người mong muốn truyền tải. Đó là lúc mà con người nghĩ đến việc thay đổi. Và đó cũng chính là lúc Semantic Web được ra đời.

Semantic Web mang sứ mệnh lớn lao trong việc thay đổi công nghệ web. Trước đây, máy tính chỉ đóng vai trò là trung tâm “chứa đựng” và “duy trì” các trang web. Tuy nhiên, với Semantic Web, máy tính sẽ phải làm nhiều hơn thế, sẽ phải “suy nghĩ” và “sử dụng” trang web một phần nào đó thay thế cho con người. Để làm được điều này không phải dễ dàng, vì máy tính là vật vô tri vô giác. Do đó, một cách tiếp cận đơn giản để giải quyết vấn đề này, bằng cách thêm vào các trang web thông thường metadata, để các máy tính có thể “đọc” được, như một ngôn ngữ đầu vào của các máy tính.

Nhận thấy được những tiềm năng và những lợi ích to lớn của Semantic Web, chúng em đã lựa chọn đề tài này, để nghiên cứu và tìm hiểu sâu hơn về Semantic Web, góp phần đem những kiến thức tìm hiểu được xây dựng thành một luận văn mang tính đóng góp cao. Trong quá trình nghiên cứu, tuy gặp phải những khái niệm và công nghệ hoàn toàn mới lạ và ít được công bố, chúng em vẫn cố gắng tìm hiểu bằng mọi cách. Lĩnh vực Semantic Web là rất rộng lớn, với một khoảng thời gian có hạn, chúng em chỉ có thể tìm hiểu được những vấn đề được coi là cơ bản và tất yếu nhất của Semantic Web. Dù vậy, chúng em rất hài lòng và tự tin với những gì tìm hiểu và nghiên cứu được sẽ mang lại nhiều lợi ích, đóng góp vào công cuộc nghiên cứu khoa học chung...

Lời cảm ơn

Đầu tiên, chúng em xin chân thành cảm ơn Khoa Mạng máy tính và Truyền thông, trường Đại Học Công Nghệ Thông Tin, Đại Học Quốc Gia TP.HCM đã tạo điều kiện cho chúng em hoàn thành tốt khoá luận này.

Chúng em xin chân thành cảm ơn Thầy Nguyễn Anh Tuấn, đã tận tình hướng dẫn, dạy dỗ, chỉ bảo chúng em từ những ngày đầu định hình khoá luận cho đến khi hoàn thành. Nhờ sự tận tình của thầy, chúng em đã hoàn thành tốt khoá luận này, bên cạnh đó cũng học hỏi được nhiều kiến thức quý báu từ thầy.

Chúng em xin chân thành cảm ơn quý Thầy Cô trong Khoa Mạng máy tính và truyền thông, trong những năm qua đã không quản ngại mệt mỏi, tận tình giảng dạy, trang bị cho chúng em những kiến thức cần thiết để hoàn thành khoá luận.

Chúng em xin ghi nhớ công ơn sinh thành dưỡng dục của cha mẹ, sự giúp đỡ của các anh, chị, bạn bè trong những năm học, cũng như sự an ủi, động viên trong những lúc khó khăn, vất vả. Dù chúng em đã dùng tất cả nỗ lực của bản thân để hoàn thành tốt khoá luận này, tuy nhiên không thể tránh khỏi những sai sót, thiếu sót, kính mong quý Thầy Cô tận tình chỉ bảo. Một lần nữa, chúng em xin chân thành cảm ơn và mong nhận được nhiều tình cảm chân thành của tất cả mọi người.

Thành phố Hồ Chí Minh, ngày ... tháng ... năm 2015

Sinh viên thực hiện khoá luận

Đặng Lê Bảo Chương và Nguyễn Bảo Duy

Mục lục

Danh sách hình vẽ

Danh sách bảng

Các từ viết tắt

KB	K nowledge B ase
KR	K nowledge R epresentation
DL	D escription L ogic
MUPS	M inimal U nsatisfiability P reserving S ub- T Boxes
HST	H itting S et T ree
HS	H itting S et

Ký hiệu

\models	models of	có nghĩa trong (KB)
$\not\models$	not a model of	không có nghĩa trong KB
\subseteq	is a subset of	là tập con của
\cap	intersect	giao với
$\neg A$	complement of A of	không phải A
$\exists R.E$	e.g <i>has</i> some E	
$\forall R.E$	e.g <i>has</i> only E	
\equiv	is equivalent to	tương đương với
\emptyset	empty set	tập hợp rỗng
\in	is member of	thuộc
\Leftarrow	preferred for left implication	
\setminus	except	ngoại trừ

For/Dedicated to/To my...

Chương 1

Giới thiệu

1.1 Tên đề tài

1.2 Nội dung và giới hạn đề tài

1.2.1 Nội dung đề tài

OWL (Web Ontology Language) là một dạng ngôn ngữ biểu diễn tri thức. Ngôn ngữ này thường được sử dụng phổ biến trong Semantic Web, và được trình bày dưới dạng RDF/XML. Ngày 27 tháng 10 năm 2009, tổ chức W3C (World Wide Web Consortium) cho công bố OWL 2, với trình chỉnh sửa Protégé và các bộ reasoner như Pellet, HermiT, v.v .

OWL và Semantic Web hiện đang được nghiên cứu và phát triển, nhằm nhanh chóng đưa vào sử dụng, vì những lợi ích rất đáng kể, được xem là phiên bản web 3.0. Do đó, nhóm quyết định nghiên cứu về OWL, về phương thức hoạt động của các bộ reasoner, và đặc biệt là thiết kế một trình chỉnh sửa OWL trên web với giao diện thân thiện và dễ sử dụng, với các tính năng gần như đầy đủ so với chương trình Protégé.

Chắc chắn trong vài năm sắp tới, Semantic Web sẽ phát triển ngày càng lớn mạnh hơn, dần dần thay đổi phương thức tiếp cận và lưu trữ dữ liệu trên web. Vậy nên, việc tìm hiểu và nghiên cứu về ngôn ngữ OWL - một trong những thành phần quan trọng của Semantic Web, có thể coi như một bước “đón đầu công nghệ”, nhằm mục đích sẵn sàng thích nghi với sự chuyển biến không ngừng của thế giới công nghệ thông tin.

Đề tài sẽ làm những việc sau:

1. Tìm hiểu về Semantic Web và Open World Assumption.

2. Tìm hiểu về ngôn ngữ Web Ontology Language (OWL) và Semantic Web Rule Language (SWRL).
3. Tìm hiểu về OWLAPI và SWRL API.
4. Tìm hiểu về nguyên lý hoạt động của OWL Reasoner (cụ thể là Pellet Reasoner).
5. Tìm hiểu về Vaadin Framework.
6. Sử dụng Vaadin Framework để xây dựng công cụ phục vụ phát triển Ontology trên web.
7. Giới thiệu những đặc điểm và tính năng nổi bật của phần mềm.
8. Kết luận và hướng phát triển nghiên cứu.

1.2.2 Giới hạn của đề tài

Lĩnh vực Semantic Web là rất rộng lớn, nhóm chỉ tập trung nghiên cứu về OWL, về OWL API [?] và SWRL API [?] để hiện thực chương trình chỉnh sửa, bên cạnh đó nghiên cứu hoạt động của Reasoner để hiện thực quá trình phân loại tự động, tìm hiểu về Vaadin Framework để xây dựng chương trình chỉnh sửa. Các vấn đề khác nằm ngoài tầm vóc của luận văn này.

1.2.3 Cấu trúc luận văn

Luận văn được chia thành các chương như sau : (cần thêm vào)

Chương 2

Giới thiệu về Open World Assumption và Semantic Web

2.1 Open World Assumption

Trước khi bắt đầu giới thiệu với sâu hơn về Ontology Web Language (OWL), chúng em xin được giới thiệu qua về giả định Thế Giới Mở (Open World Assumption - OWA)[?] được Semantic Web chấp nhận và phân biệt giả định này với giả định Thế Giới Đóng (Closed World Assumption - CWA).

Closed World Assumption Giả định Thế Giới Đóng (CWA) là giả định mà những điều không chắc hoặc không có cơ sở để chứng minh là **đúng** sẽ được chấp nhận là **sai**.

Open World Assumption Giả định Thế Giới Mở (OWA) thì ngược lại, với những điều không chắc hoặc không có cơ sở để chứng minh là **đúng** sẽ được chấp nhận là **chưa biết**.

Ví dụ Xem xét một câu nói sau đây: "A là một công dân của nước Hoa Kỳ". Nếu có ai đó hỏi "A có phải là một công dân của Việt Nam hay không?". Xét theo CWA, câu trả lời là *không*, ngược lại với OWA thì câu trả lời là *chưa biết*.

2.1.1 Vậy OWA và CWA được sử dụng khi nào ?

Giả định thế giới đóng (CWA) được sử dụng khi một hệ thống đã có đầy đủ thông tin. Đây là trường hợp được áp dụng cho nhiều ứng dụng cơ sở dữ liệu. Ví dụ, xem xét một tình huống một ứng dụng cơ sở dữ liệu đặt vé máy bay, chúng ta tìm kiếm đường bay

thăng Phú Quốc và Hà Nội, và kết quả là nó không tồn tại trong cơ sở dữ liệu (không quan tâm đến thực tế có hay không có đường bay này). Và theo CWA nên câu trả lời từ cơ sở dữ liệu là : "Không có đường bay thăng Hà Nội - Phú Quốc" (Một giả định là thực tế cũng không tồn tại đường bay này do cơ sở dữ liệu không biết). Đây là dạng ứng dụng mà người dùng mong đợi một câu trả lời chính xác (phổ biến ở các cơ sở dữ liệu quan hệ).

Ngược lại với Giả định thế giới đóng, Giả định thế giới mở được áp dụng trên một hệ thống mà thông tin được cung cấp không đầy đủ. Đây là trường hợp chúng ta một biểu dạng một dạng tri thức (a.k.a Ontologies) và chúng ta muốn khám phá những thông tin mới tiềm ẩn trong đó. Ví dụ, xem xét một hệ thống lưu trữ tiền sử bệnh lý của bệnh nhân. Nếu cơ sở dữ liệu không chứa thông tin về một dạng dị ứng cụ thể mà bệnh nhân mắc phải, điều đó không đồng nghĩa là bệnh nhân đó không mắc phải nó trên thực tế. Từ đó câu trả lời từ cơ sở dữ liệu theo chuẩn OWA sẽ là : "Không rõ bệnh nhân này có mắc phải dị ứng đó không, trừ khi những thông tin đầy đủ hơn được cung cấp".

2.1.2 So sánh OWA và CWA

Giả định Thế Giới Đóng không chỉ là trả về các câu trả lời "*không*" và Giả định Thế Giới Mở không chỉ là trả về "*không biết*". Lấy lại ví dụ trên: "**A** là một công dân Hoa Kỳ" và giả sử khẳng định sau là đúng: "Một người chỉ có thể là công dân của một quốc gia". Chúng ta cùng xem tiếp câu sau: "A là công dân Việt Nam". Xét theo Giả định Thế Giới Đóng (CWA), phát biểu này có thể gây ra lỗi, vì nó mâu thuẫn với 2 phát biểu ban đầu giả sử chúng ta biết Mỹ và Việt Nam là 2 quốc gia khác nhau . Nếu đem phát biểu vừa rồi xét trong Giả định thế giới mở, thay vì gây lỗi, nó sẽ suy ra một phát biểu logic như sau: "Nếu một người chỉ có thể là công dân của một quốc gia, và nếu A là công dân của Hoa Kỳ và Việt Nam, thì Hoa Kỳ và Việt Nam phải cùng là một quốc gia".

Lưu ý trong ví dụ, trường hợp CWA chúng ta đã giả sử Hoa Kỳ và Việt Nam là những quốc gia khác nhau, thật ra đây chính là Giả Định tên duy nhất (Unique Named Assumption) *. Mặc định các hệ thống OWA không áp dụng UNA, tuy vậy chúng ta hoàn toàn có thể áp dụng UNA cho OWA. Trong ví dụ trên, giả sử nếu chúng ta biết hết tất cả tên của các quốc gia trên thế giới và khẳng định rõ là mỗi tên đại diện cho duy nhất một nước hay ví dụ trên sẽ có thêm phát biểu là "Hoa Kỳ và Việt Nam là 2 quốc gia khác nhau". Kết quả dẫn đến là toàn bộ phát biểu trong ví dụ trên sẽ thiếu nhất quán (inconsistent) **

* *Unique Name Assumption*: http://en.wikipedia.org/wiki/Unique_name_assumption

** *Tính thiếu nhất quán của ontology sẽ được đề cập ở các chương tiếp theo*

2.1.3 Bảng tổng hợp các đặc điểm của OWA và CWA [?]

[?].

Cách tiếp cận theo hướng Cơ sở dữ liệu quan hệ (CWA)	Cách tiếp cận theo hướng Semantic Web (OWA)
Hệ thống mà những cái chưa biết là đúng sẽ được xem là sai. Mọi thứ bị cấm cho đến khi nó được cho phép	Hệ thống mà cái chưa có cơ sở khẳng định là đúng-sai sẽ được xem là chưa biết. Mọi thứ được cho phép cho đến khi nó bị cấm.
Giả định tên duy nhất (UNA)	Những tên/nhãn trùng nhau được cho phép
Giả định tên duy nhất quy định mỗi tên đại diện cho những cá thể khác nhau trong thế giới	OWL cho phép sử dụng các nhãn khác nhau nhưng đồng nghĩa để đại diện cho cùng một đối tượng hoặc một tên có thể dùng chỉ nhiều đối tượng khác nhau. Các khẳng định về nhận dạng phải được khai báo một cách rõ ràng.
Thông tin được đầy đủ	Thông tin không đầy đủ
Dữ liệu tồn tại trong hệ thống được giả sử đã được hoàn chỉnh (những thông tin thiếu thường được xử lý bằng giá trị null trong SQL*, nhưng có thể gây mâu thuẫn về tính đúng đắn của chính nó). Đây còn được gọi là giả định trong một miền đóng**	Một nguyên lý tối thiểu của OWA chính là sự không đầy đủ của thông tin. Một mệnh đề hiển nhiên đúng khi nói rằng các thuộc tính của một đối tượng hay cá thể cụ thể có thể còn thiếu hoặc chỉ mới được biết đến một phần.
Single Schema(Một bối cảnh đơn lẻ)	Biểu diễn được trên nhiều bối cảnh/thế giới khác nhau
Single Schema là cần thiết để xác định rõ phạm vi và cách lý giải trong một miền tri thức hữu hạn	Các khai báo về bối cảnh (schema) và các cá thể dữ liệu được tách ra. Nên sẽ có thể nhiều cách giải thích (trong nhiều miền tri thức) cho cùng một dữ liệu.
Những ràng buộc toàn vẹn (Integrity Constraints)	Những tiên đề logic (Logical Axioms)

Những ràng buộc toàn vẹn nhằm ngăn các giá trị "*Không hợp lệ*" không được khai báo trong mô hình quan hệ. Điều này rất hữu dụng trong việc kiểm tra tính hợp lệ và cú pháp của dữ liệu input. Những ràng buộc về số lượng nghiêm ngặt được sử dụng để kiểm tra tính hợp lệ của dữ liệu. Ví dụ: char(40) -> quy định một column chỉ chứa tối đa 40 kí tự.

Những tiên đề logic cho phép tạo ra những hạn chế thông qua miền và vùng giới hạn mà các đặc tính của đối tượng hướng đến. Tất cả mọi thứ đều có thể đúng trừ khi chúng bị chứng minh là sai, và tồn tại nhiều mô hình (Knowledge Base) thỏa tiên đề. Nhờ vậy, OWA sở hữu khả năng suy luận mạnh mẽ, mặc dù chức năng này vẫn còn kém trực quan ở thời điểm hiện tại. Những hạn chế về số lượng và phạm vi dữ liệu có cách thể hiện khác nhau dành cho đối tượng (được suy luận) hoặc dạng dữ liệu.

Logic không đơn điệu (Non-monotonic Logic)

Tập hợp các kết luận đảm bảo dựa trên nền tảng của một cơ sở tri thức cho trước không thì sẽ không tăng (trên thực tế chúng có vẻ giảm đi) so với kích thước của cơ sở tri thức

Logic đơn điệu

Các giả thiết về bất kì sự thực hiển nhiên nào được có thể được mở rộng bằng cách thêm vào những giả định mới. Những khẳng định mới thêm vào có xu hướng làm giảm sự suy luận và hàm ý có thể được áp dụng. Một mảng tri thức mới nào đều không làm giảm đi những cái đã biết. Những tri thức mới có thể nổi lên thông qua việc suy luận.

Cố định và khó chỉnh sửa

Thay đổi bối cảnh (schema) sử dụng đòi hỏi phải thiết kế lại kiến trúc cơ sở dữ liệu, không có khả năng tự mở rộng

Có khả năng tái sử dụng và mở rộng

Được thiết kế từ nền tảng nhằm hướng đến việc tái sử dụng những ontologies (tập hợp các tiên đề hay phát biểu hiển nhiên đúng) có sẵn và có khả năng mở rộng. Việc thiết kế cơ sở dữ liệu và quản lý có thể linh động hơn, với bối cảnh phát triển không ngừng.

Cấu trúc phẳng (table); Phân loại dữ liệu rõ ràng (Strong typing)

Cấu trúc đồ thị (Graph); Phân loại dữ liệu mở

Thông tin được tổ chức thành các bảng phẳng; các mối quan hệ và kết nối giữa các bảng dựa trên khóa ngoại hoặc các kết nối giữa nhiều bảng với nhau (JOIN trong SQL). Định hướng phân loại các dạng dữ liệu rõ ràng (char, string, smallint, int, bigint, etc.)

Các công cụ phát triển, lưu trữ và truy vấn dữ liệu

Ngôn ngữ SQL và các câu truy vấn đã được phát triển rất tối ưu. Nhiều phần mềm hỗ trợ rất tốt cho việc phát triển. Không hỗ trợ disjunction (phép logic OR) - không tồn tại 1 record nào thuộc 2 bảng khác nhau; tính phủ định phải được thực hiện theo nguyên tắc Negation As Failure (NAF). Việc tính tổng và thống kê dễ dàng hơn nhờ vào UNA. Câu trả lời đóng (một câu trả lời được truyền cho tác vụ tính toán kế tiếp) dễ hơn so với OWA

Thông tin được tổ chức dạng đồ thị (RDF Graph), hỗ trợ phân tích các mối quan hệ và kết nối. Ngược lại với CWA, dạng dữ liệu (Datatypes) rất linh động (String), ngoài ra cũng cho phép thêm vào các phát biểu định nghĩa về dữ liệu để hỗ trợ phân loại dữ liệu rõ ràng hơn. Datatype được xem như các lớp (Classes), và giá trị của chúng được xem như những định danh riêng biệt (nói cách khác giá trị dữ liệu được xem giống như một đối tượng).

Các công cụ phát triển, lưu trữ và truy vấn dữ liệu

Ngôn ngữ SPARQL và những ngôn ngữ Rule mới được phát triển để phục vụ cho việc truy vấn (SQWRL, Jena Rule, etc.); hiệu năng khi mở rộng ở quy mô lớn vẫn còn đang được cải thiện. Các câu truy vấn đòi hỏi phải có thông tin ngữ cảnh để chọn đúng tập các dữ liệu muốn xem. Tính phủ định và logic OR (disjunction) được cho phép và được xây dựng mạnh hơn. Vd: có thể khai báo một Class C là Disjunction $A \cup B$. Chưa có nhiều công cụ hỗ trợ được phát triển.

Kết luận OWA được áp dụng lên những hệ thống mà thông tin chưa được cung cấp đầy đủ. Và Web chính là một hệ thống với những thông tin còn thiếu. Sự thiếu vắng thông tin trên Web có nghĩa là sẽ có những thông tin không được làm rõ. Điều này giải thích tại sao Semantic Web sử dụng OWA, vì nhờ đó Semantic Web có thể suy ra được những thông tin mới.

* Null statement in SQL: http://en.wikipedia.org/wiki/Null_%28SQL%29

** Domain-closure Assumption: http://en.wikipedia.org/wiki/Integrally_closed_domain

2.2 Web ngữ nghĩa (Semantic Web)

2.2.1 Giới thiệu

Semantic Web được các nhà nghiên cứu kỳ vọng sẽ trở thành Web 3.0, với đặc trưng riêng biệt là phương thức liên kết dữ liệu (linked data) giữa các hệ thống hoặc các thực thể cho phép thể hiện được nhiều hơn, và rõ ràng hơn mối liên kết giữa các dữ liệu trên mạng lưới web toàn cầu. Cụ thể hơn, Semantic Web có khả năng chuyển đổi văn bản HTML của con người (human readable HTML documents) sang ngôn ngữ máy tính (machine readable documents), giúp cho máy tính làm được nhiều công việc suy nghĩ hơn cho con người [?].

Ngày nay, đa phần dữ liệu trên web được cung cấp dưới dạng trang web (web pages) - văn bản HTML được liên kết với nhau bằng các liên kết (hyperlinks). Cả người và máy tính đều có thể dễ dàng đọc hiểu những văn bản đó, tuy nhiên thay vì tìm kiếm những từ khoá trong trang web, máy tính lại gặp trở ngại khi chọn lọc những ý nghĩa trong các văn bản đó. Một trang web chứa rất nhiều thông tin, nhưng những thông tin đó không phải là những thông tin thô - mà chỉ là những văn bản HTML được xây dựng từ cơ sở dữ liệu.

- Chuyển những trang web dữ liệu thành những tiến trình xử lý thông minh nhân tạo (giúp trang web phải “suy nghĩ” để xử lý giúp con người).
- Khuyến khích các công ty, các doanh nghiệp và các cá nhân trình bày dữ liệu tự do hơn, theo một quy chuẩn mở.
- Khuyến khích các doanh nghiệp sử dụng dữ liệu đã có sẵn trên web.

2.2.2 Semantic Web dựa trên giả định thế giới mở (OWA)

Như chúng ta đã biết thì về mục tiêu mà Semantic Web hướng đến, để đạt được những mục tiêu đó, cần có khả năng xử lý thông tin ở khắp mọi nơi đòi hỏi các tiêu chuẩn cũng như các nguyên lý tổ chức dữ liệu sẽ không giống như trước (khi chúng ta vẫn còn tổ chức dữ liệu thành các bảng dữ liệu quan hệ) theo giả định thế giới đóng. Do đó với tính chất muốn bao quát hết tất cả thông tin trên web (gồm luôn những thông tin chưa đầy đủ - incomplete information) thì Semantic Web đã lấy Giả định Thế Giới Mở đã được đề cập ở phần trên nhằm đảm bảo một hệ thống luôn sẵn sàng mở rộng và tiếp nhận thông tin mới mà không đòi hỏi phải thiết kế lại.

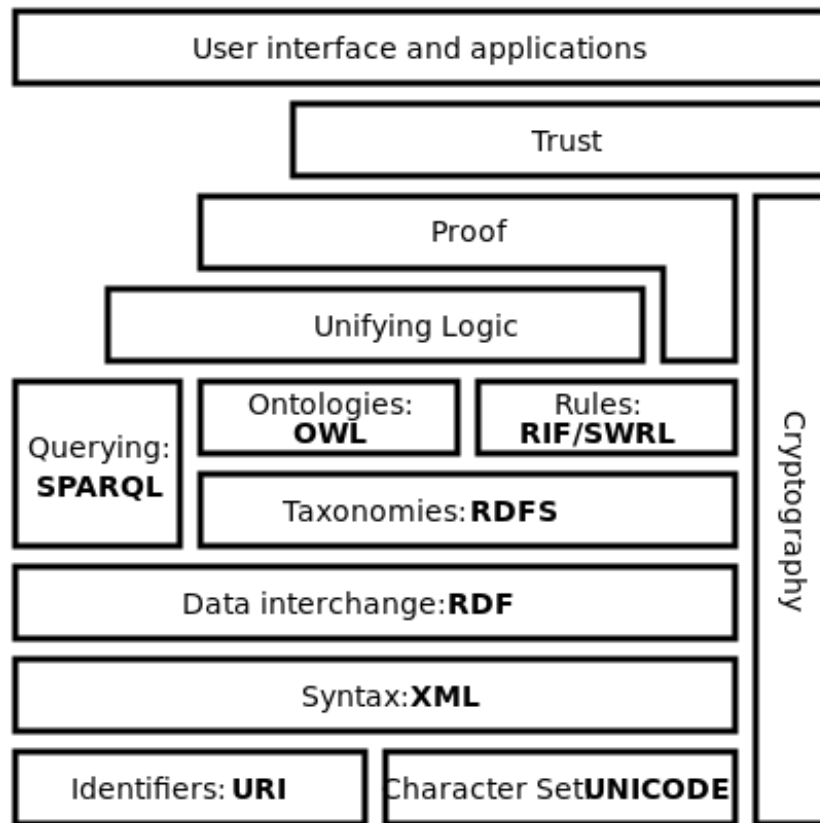
2.2.2.1 Các tiêu chuẩn và thành phần của Semantic Web

Khái niệm "Semantic Web" thường được sử dụng cụ thể hơn nhằm chỉ đến những định dạng và công nghệ để hiện thực hóa nó. Việc tổ chức, tập hợp và phục hồi dữ liệu liên kết thực hiện được nhờ vào các công đặc tả chính thức về các khái niệm, định nghĩa và mối quan hệ trong một vùng tri thức (knowledge domain) cho trước. Tất cả các công nghệ này đều được quy định thành một tiêu chuẩn của W3C [?]. Các tiêu chuẩn được liệt kê dưới đây

- [Resource Description Framework](#), một phương thức chung để biểu diễn thông tin cho semantic web.
- [RDF Schema](#)
- [Simple Knowledge Organization System](#) (SKOS)
- [SPARQL](#) - Ngôn ngữ truy vấn dữ liệu biểu diễn dưới dạng RDF.
- [Notation3](#), thiết kế với tiêu chí hiểu được bởi con người.
- [N-Triples](#), một định dạng dùng để lưu và truyền dữ liệu.
- [Turtle](#) (Terse RDF Triple Language)
- [Web Ontology Language](#) (OWL), một họ các ngôn ngữ biểu diễn tri thức
- Rule Interchange Format (RIF), một framework chung của các ngôn ngữ điều luật web hỗ trợ chuyển đổi nhiều điều luật khác nhau trên web

Hình Semantic Web Stack[?] miêu tả kiến trúc của Semantic Web. Chắc rằng mà mối quan hệ giữa các thành phần được tổng hợp lại dưới đây:

- XML cung cấp một cú pháp cơ bản nhất cho nội dung bên trong tài liệu, và không có liên quan ngữ nghĩa gì đến nội dung ngữ nghĩa mà nội dung nó chứa. XML không phải là một thành phần cần thiết trong các công nghệ Semantic Web trong hầu hết các trường hợp, tồn tại cú pháp thay thế khác như Turtle *.
- XML Schema là một ngôn ngữ dùng để cung cấp và hạn chế cấu trúc nội dung của các thành phần nằm trong tài liệu XML, nói cách khác nó giúp chúng ta danh nói dung mà tài liệu đó chứa là gì. Ví dụ: OWL/XML vs. RDF/XML
- RDF [?] là một ngôn ngữ đơn giản dùng để diễn tả các mô hình dữ liệu (ở đây muốn chỉ đến các nguồn dữ liệu web) và mối quan hệ của chúng. Một mô hình dữ liệu theo RDF có thể được biểu diễn bằng nhiều cú pháp khác nhau, vd: RDF/XML,



HÌNH 2.1: The Semantic Web Stack

N3, Turtle và RDFa. Có thể nói RDF chính là thành phần cơ bản và quan trọng nhất của Semantic Web.

- RDF Schema [?] mở rộng RDF và là từ vựng để đặc tả các thuộc tính và lớp trong các tài nguyên dựa trên RDF, với ngữ nghĩa dựa trên các việc tạo ra nhiều phân cấp lớp và thuộc tính.
- OWL thêm nhiều từ vựng hơn để diễn các thuộc tính và lớp, và điểm quan trọng là nó thêm các từ vựng để đặc tả mối quan hệ giữa các lớp với nhau, vd: ranh giới riêng biệt giữa các lớp với nhau (disjointness), các quy định với số lượng (cardinality), cung cấp nhiều loại dữ liệu cho các thuộc tính, và các đặc tính của các thuộc tính (vd: đối xứng/ bất đối xứng, và các lớp liệt kê).
- SPARQL là một giao thức và ngôn ngữ truy vấn dữ liệu dành cho tài nguyên của semantic web.
- RIF (W3C Rule Interchange Format) là một ngữ ngữ XML để biểu diễn điều luật web mà máy tính có thể thực thi.

* Turtle: [http://en.wikipedia.org/wiki/Turtle_\(syntax\)](http://en.wikipedia.org/wiki/Turtle_(syntax))

Kết luận Trên đây chúng em chỉ liệt kê những thành phần và tiêu chuẩn cơ bản nhất mà tổ chức W3C đã vạch ra nhằm xây dựng một mô hình Web ngữ nghĩa của tương lai. Nội dung đề tài của chúng em chỉ hạn chế trong việc nghiên cứu và khai thác ngôn ngữ ontology web nhằm khai thác tiềm năng về mặt ngữ nghĩa (suy luận ra những thông tin mới dựa trên những suy luận từ ngữ nghĩa của những thông tin được khai báo) nhằm phục vụ cho việc phân loại thông qua các thuộc tính của sản phẩm. Chương kế tiếp sẽ đi qua tìm hiểu về Ontology Web Language(OWL) và Semantic Web Rule Language (SWRL) hai thành phần chính giúp hình thành khả năng phân loại tự động của đề tài này.

Chương 3

Chi tiết Ontology Web Language

Giới thiệu - Như đã được đề cập trong phần cuối của chương trước, chức năng chính của OWL là một ngôn ngữ ontology cung cấp ngữ nghĩa cho Semantic Web. Trong nội dung chương này, chúng em sẽ giới thiệu về cú pháp, định dạng và chi tiết các đặc tính của ngôn ngữ Ontology Web. Phiên bản Ontology Web Language chúng em sử dụng là phiên bản 2 được tổ chức W3C khuyến khích sử dụng so với phiên bản OWL 1.1 .

3.1 Khái quát về OWL 2 [?]


3.1.1 Tổng quan

Hình trên cho chúng ta cái nhìn tổng quan về các định dạng file, các loại cú pháp và cách khả năng serialization thành RDF Graph của Ontology. Như chúng ta thấy trong hình thì hình eclipse ở giữa thể hiện khái niệm trừu tượng của một ontology, có thể hiểu là một cấu trúc trừu tượng hay một đồ thị RDF. Chúng ta có thể dùng nhiều cú pháp để biểu diễn ontology và định dạng chúng dưới dạng file khác nhau (Syntax layer trong hình), các định dạng và cú pháp này hoàn toàn có thể chuyển đổi qua lại với nhau. Lớp ngữ nghĩa trong hình (semantic layer) cho thấy ngữ nghĩa được quy định theo 2 tiêu chuẩn kỹ thuật khác nhau là Direct Semantics và RDF-Based Semantics.

Phần lớn những người phát triển Ontology bằng OWL2 sẽ chỉ cần 1 cú pháp (tương đương với 1 định dạng file) và một dạng biểu diễn ngữ nghĩa.

3.1.2 Ontologies

Bất kì ontology OWL2 nào đều có thể được định dạng như một đồ thị RDF. Mối quan hệ giữa 2 cách này được quy định bởi cách tài liệu Mapping to RDF Graphs document



Figures/owl2structure.png

HÌNH 3.1: Cấu trúc của OWL 2

[[OWL 2 RDF Mapping](#)] [?], trong tài liệu này định nghĩa rất rõ ràng một bảng map từ định dạng cấu trúc của ontology qua đồ thị RDF, và ngược lại.

3.1.3 Cú pháp

Trong thực tế, một cú pháp cụ thể rất cần thiết để lưu trữ các OWL2 Ontologies và để trao đổi chúng giữa các công cụ và ứng dụng khác nhau. Cú pháp đầu tiên có khả năng hoán đổi là RDF/XML [[RDF Syntax](#)] [?]. Ngoài RDF/XML có khả năng cung cấp khả năng tương tác giữ nhiều ứng dụng OWL2 khác nhau, các loại cú pháp khác đều có thể được sử dụng. Dưới đây là bảng so sánh và liệt kê các cú pháp.

3.1.4 Một số ví dụ của các syntax

3.1.4.1 Functional Syntax

```
// Khai báo lớp
```

Tên cú pháp	Mô tả	Trạng thái	Mục đích sử dụng
RDF/XML	Mapping to RDF Graphs [?] [?]	Bắt buộc	Hoán đổi được (có thể viết và đọc được bằng nhiều phần mềm OWL2)
OWL/XML	XML Serialization [?]	Tùy chọn	Xử lý dễ dàng hơn bằng công cụ XML.
Functional Syntax	Structural Specification [?]	Tùy chọn	Dễ đọc và hiểu được.
Manchester Syntax	Manchester Syntax [?]	Tùy chọn	Có ưu thế hơn để đọc/ghi DL Ontologies
Turtle	Mapping to RDF Graphs [?]	Tùy chọn, không được công nhận chính thức	Có ưu thế để đọc/ghi RDF triples

BẢNG 3.1: Bảng so sánh các cú pháp của OWL2

```

Declaration (Class (Animal)) 1
Declaration (Class (Grass))
// Khai báo Object Property
Declaration (ObjectProperty (canEat))
// Khai báo sub Class expression
SubClassOf (Cow Animal)
// Khai báo sub Class Expression
SubClassOf (Cow ObjectSomeValueFrom(canEat Grass))

```

3.1.4.2 RDF/XML Syntax

```

T(Animal) rdf:type owl:Class
T(Grass) rdf:type owl:Class
T(canEat) rdf:type owl:ObjectProperty
T(Cow) rdfs:subClassOf T(Animal)
T(Cow) rdfs:subClassOf T(_:x owl:someValuesFrom T(Grass))

```

3.1.4.3 OWL/XML Syntax

```

<Declaration>
  <Class IRI="#Animal"/> // Khai báo lớp Animal
</Declaration>
<Declaration>
  <Class IRI="#Grass"/> // Khai báo lớp Grass
</Declaration>

```

```

<Declaration>
    <ObjectProperty IRI="#canEat"/> // Khai báo Object Property
</Declaration>
<SubClassOf>
    <Class IRI="#Cow"/>
    <Class IRI="#Animal"/> // Khai báo sub Class expression
</SubClassOf>
<SubClassOf>
    <Class IRI="#Cow"/>
<ObjectAllValuesFrom>
<ObjectProperty IRI="#canEat"/> // Khai báo sub Class expression
    <Class IRI="#Grass"/>
</ObjectAllValuesFrom>
</SubClassOf>

```

3.1.4.4 Manchester Syntax

```

Class: Cow
    SubClassOf: Animal
    SubClassOf: canEat some Grass
Class: Grass
ObjectProperty: canEat

```

3.2 Các đặc tính chi tiết của OWL2

Ngôn ngữ Ontology Web Language 2 (hay OWL 2) đảm nhiệm chức năng thể hiện ngữ nghĩa cho Semantic Web như chúng ta đã thấy trong hình Semantic Web Stack. OWL2 ontologies cung cấp các lớp (class), đặc tính (property), cá thể (individual) và các giá trị dữ liệu (data value), tất cả chúng sẽ được biểu diễn bằng các tài liệu và cú pháp được đề cập ở trên, phần này sẽ đi vào giải thích các thành phần vừa nêu ra của một ontology được viết bằng ngôn ngữ OWL2. Đây cũng chính là các thành phần cấu trúc mà thư viện OWLAPI [?] áp dụng để xây dựng nên bộ thư viện giúp người phát triển ứng dụng ontology tương tác dễ dàng hơn với các tài liệu ontology. Đầu tiên chúng em xin được giới thiệu tới đối tượng lớn nhất trong quy định của ngôn ngữ OWL2.

Tiếp theo, chúng em xin được trình bày về một qua các định nghĩa những thành phần cấu tạo của ontology mà có ảnh hưởng đến quá trình xây dựng ứng dụng, các thành phần không liên quan sẽ không được đề cập.

3.3 Định nghĩa các thành phần cấu tạo nên một ontology

3.3.1 Ontology IRI và Version IRI

Mỗi ontology đều có thể có một *ontology IRI* [?] (Internationalized Resource Identifier), dùng để định danh cho ontology. Nếu một ontology có một ontology IRI, thì ontology này có thể có thêm một version IRI, dùng để xác định phiên bản cho ontology này. Version IRI có thể trùng hoặc không cần thiết phải trùng với ontology IRI. Một ontology không có ontology IRI thì không có version IRI. Dưới đây là những quy ước chọn ontology IRIs và version IRIs trong OWL2. Những đặc điểm kỹ thuật này không cung cấp cơ chế nào để làm chúng phải được tuân theo trên toàn hệ thống web. Tuy nhiên, những công cụ hay ứng dụng OWL2 nên sử dụng những quy ước này để dễ dàng tìm ra lỗi trong những ontology mà chúng xử lý.

- Nếu một ontology có một ontology IRI nhưng không có version IRI, thì *không nên tồn tại* một ontology với trùng ontology IRI vừa đặt.
- Nếu một ontology có một ontology IRI và một version IRI, thì *không nên tồn tại* một ontology khác với trùng ontology IRI và version IRI vừa đặt.
- Tất cả các cách kết hợp khác của ontology IRI và version IRI không cần đòi hỏi tính duy nhất (unique). Như vậy 2 ontologies khác nhau có thể không có ontology IRI và version IRI; tương tự, một ontology chưa một ontology IRI có thể cùng tồn tại cùng với một ontology khác có cùng ontology IRI vừa đặt **và** các version IRI của các ontologies này **phải** khác nhau.

Ontology IRI và các version IRI kết hợp với nhau giúp định danh một phiên bản cụ thể của của ontology từ một bộ chứa tất cả các phiên bản của một ontology cụ thể nào đó được định danh chung bằng ontology IRI. Trong mỗi bộ ontology như vậy, sẽ có chính xác một ontology được dùng nhưng một ontology hiện hành - khi dùng ontology IRI để truy vấn ontology mà không đề cập version IRI mặc định ontology có version IRI hiện hành sẽ được trả về.

3.3.2 Thực thể, trực nghĩa và cá thể ẩn danh - Entities, Literals and Anonymous Individuals

Các thực thể (entities) là thành phần cơ bản nhất của OWL2 Ontology, chúng định nghĩa các từ vựng - cụ thể là những đặt tên ra các khái niệm (named term) - của một ontology. Bên cạnh các thực thể, OWL 2 ontologies thường có thêm các trực nghĩa



HÌNH 3.2: Entities, Literals, Anonymous Individuals trong OWL2

(literals), như strings hay integers.

Cấu trúc của các thực thể và trực nghĩa trong OWL 2 được thể hiện trong hình bên. Các lớp (classes), kiểu dữ liệu (datatypes), đặc tính đối tượng (object properties), đặc tính dữ liệu (data properties), thuộc tính chú thích và các cá thể có tên đều được gọi chung là các thực thể (entities), tất cả chúng được định danh bằng một IRI duy nhất.

- Lớp (class) đại diện cho một tập gồm nhiều cá thể (individuals).
- Kiểu dữ liệu (datatype) là một tập của các trực nghĩa như strings hoặc integers.
- Đặc tính đối tượng và dữ liệu (object & data property) được sử dụng để biểu diễn các mối quan hệ giữa các cá thể với cá thể khác và giữa cá thể với kiểu dữ liệu trong một miền (domain) nào đó.

- Đặc tính chú thích (annotation) được dùng để đưa thêm những thông tin không có tính ngữ nghĩa (non-logical) như các chú thích, giải nghĩa, ngôn ngữ gắn với ontologies, các phát biểu/ tiên đề (axioms) và các thực thể.
- Các cá thể có tên có thể được dùng để biểu diễn một đối tượng cụ thể từ một lớp nào đó.

Bên cạnh các cá thể có tên, OWL2 còn cung cấp một khái niệm gọi là các cá thể ẩn danh (anonymous individuals) - là cá thể tương tự với các node trống (blank nodes) trong RDF Concept [?] và được truy xuất ngay bên trong ontology mà chúng được sử dụng. Cuối cùng, OWL 2 cung cấp thêm cho các trực nghĩa (literals), một dạng dữ liệu string gọi là định dạng nghĩa bằng từ ngữ (lexical form) và một dạng dữ liệu để chỉ dẫn cách ontology có thể hiểu chuỗi này.

3.3.2.1 Lớp (Classes)

Lớp được hiểu như tập hợp các cá thể. Hai lớp với IRIs *owl:Nothing* và *owl:Thing* là các lớp được định nghĩa sẵn trong OWL2 với ý nghĩa như sau:

- **owl:Thing** là tập hợp gồm tất cả các cá thể.
- **owl:Nothing** là tập hợp rỗng,

Không nên sử dụng 2 định nghĩa trên để gán cho bất kì lớp nào trong OWL 2 DL Ontology.

3.3.2.2 Kiểu dữ liệu (datatypes)

Kiểu dữ liệu là thực thể được xem như tập hợp của các giá trị dữ liệu. Như vậy, kiểu dữ liệu cũng tương tự lớp, khác biệt chính là thay vì chứa các cá thể (individuals) như lớp thì lại chứa các giá trị dữ liệu như strings, numbers, Kiểu dữ liệu có thể được dùng tạo ra các dữ liệu giới hạn (datarange).

3.3.2.3 Đặc tính đối tượng (object properties)

Đặc tính đối tượng kết nối các cặp cá thể - tạo ra mối liên hệ (relationship) giữa các cá thể. Tương tự lớp cũng có 2 đặc tính đối tượng được định nghĩa sẵn trong OWL 2 với ý nghĩa như sau:

- **owl:topObjectProperty** kết nối tất cả các cặp cá thể có thể kết nối.
- **owl:bottomObjectProperty** không kết nối bất kì cặp cá thể nào.

Cũng không nên sử dụng 2 định nghĩa trên để gán cho bất kỳ đặc tính đối tượng nào trong OWL 2 DL Ontology

3.3.2.4 Đặc tính dữ liệu (data properties)

Đặc tính dữ liệu liên kết các cá thể với các trực nghĩa. Trong một vài hệ thống biểu diễn tri thức, đặc tính dữ liệu chức năng được gọi là thuộc tính. Hai định nghĩa sẵn *owl:topDataProperty* và *owl:bottomDataProperty* có ý nghĩa như sau

- **owl:topDataProperty** liên kết tất cả cá thể với tất cả các trực nghĩa.
- **owl:bottomDataProperty** không liên kết bất kì cá thể với trực nghĩa nào.

Chương 4

Các nguyên nhân dẫn đến tính thiếu nhất quán trong ontology

4.1 Các định nghĩa cần lưu ý [?]

Unsatisfiable Class/Concept dùng để chỉ một lớp hay một khái niệm trong một ontology mà ngữ nghĩa xung đột với ngữ nghĩa khác được nêu ra trong ontology hay có thể nói là các phát biểu về lớp hay khái niệm này mâu thuẫn với nhau hoặc mâu thuẫn với những phát biểu khác trong ontology.

Ví dụ Cow

```
SubClassOf: Vegetarian
Vegetarian
SubClassOf: Animal and eats only Plant
DisjointClasses:
Plant, Animal
```

Giải thích Trong ví dụ trên thì MadCow chính là một lớp không hợp lý do trong các phát biểu logic của nó mâu thuẫn với nhau Cow là lớp con của Vegetarian mà Vegetarian chỉ ăn Plant (nghĩa là ngoài Plant, Vegetarian không ăn thứ gì khác) trong khi đó khai báo của lớp MadCow là lớp con của Cow và ăn một số Sheep (Sheep là một lớp con của Animal).

Từ đó việc lý luận có thể đưa ra giả định sai là Sheep cũng có khả năng là một phần của Plant . Điểm quan trọng là Plant và Animal là 2 DisjointClasses, nói cách khác không tồn tại một cá thể nào vừa thuộc lớp Plant và vừa thuộc lớp Animal. Như vậy trong tất cả các phát biểu logic ở ví dụ trên đã có 2 phát biểu

gây mâu thuẫn chính là `eats only Plant` và `eats some Sheep`, và chúng làm cho lớp `MadCow` trở nên bất hợp lý (*unsatisfiable*).

Incoherent Ontology dùng để chỉ một *ontology/model* có ý nghĩa không mạch lạc rõ ràng do nó có chứa ít nhất một *Unsatisfiable Class/Concept* và với điều kiện là trong những *Unsatisfiable Class* này không được chứa bất kì một cá thể (*Individual*) nào.

Giả sử ta có ontology A chứa các phát biểu trong ví dụ trên ngoại trừ phát biểu cuối cùng `Individual: Dora type: MadCow` thì ta có thể nói ontology A không mạch lạc rõ ràng do nó chứa *unsatisfiable class* là `MadCow`. Chúng ta vẫn có sử dụng được ontology A vì nó vẫn còn tính nhất quán (*Consistency*) miễn là không có phần tử nào thuộc lớp `MadCow`.

Inconsistent Ontology dùng để chỉ một ontology chứa ít nhất một *Unsatisfiable Class* và có ít nhất một cá thể (*Individual*) thuộc một trong những lớp *unsatisfiable* này. Như đã thể hiện trong ví dụ đầu tiên thì cá thể `Dora` thuộc lớp `MadCow` (Một lớp *unsatisfiable* thì không nên phép có bất kì cá thể nào nếu như chúng ta muốn đảm bảo tính *consistency* cho ontology), như vậy bất kì ontology nào có những phát biểu trên đều được coi là không nhất quán (*inconsistency*), điều này đồng nghĩa là ontology đó không thể sử dụng được nữa.

4.2 Các nguyên nhân phổ biến dẫn đến tính thiếu nhất quán (Inconsistency)[?]

Các nguyên nhân dẫn đến tính thiếu nhất quán trong ontology gây bởi các lỗi được phân loại thành lỗi gây ra bởi phát biểu ở mức độ lớp (Class level - TBox), các lỗi gây ra bởi phát biểu ở mức độ cá thể (Instance/Individual level - ABox) và lỗi gây ra bởi sự kết hợp của cả 2 nguyên nhân vừa nêu trên.

4.2.1 Khởi tạo cá thể cho một Unsatisfiable Class) - (TBox + ABox)

- Khởi tạo cá thể cho một *Unsatisfiable Class* được xem là nguyên nhân phổ biến nhất gây ra tính thiếu nhất quán trong ontology.
- Ví dụ:

`Individual: Dora type: MadCow`

- Chúng ta không quan tâm đâu là nguyên nhân làm cho MadCow trở nên mâu thuẫn, chỉ cần biết là một Unsatisfiable Class thì không nên có bất kì cá thể nào trong đó. Rõ ràng là không có bất kì ontology nào mà cá thể Dora có thể đáp ứng các điều kiện như trong ví dụ đầu tiên, nói cách khác không tồn tại model nào có thể thỏa được điều kiện trên. Chúng ta phát biểu đó là một ontology không nhất quán.

4.2.2 Khởi tạo cá thể thuộc 2 class được disjoint với nhau (TBox + ABox)

- Đây là một trường hợp dễ bắt gặp vì nó sai ngay trong phát biểu về logic.
- Ví dụ

Individual: Dora

Types: Vegetarian, Carnivore

DisjointClasses: Vegetarian, Carnivore

- Lớp A disjoint với B khi và chỉ khi lớp A không có chung bất kì một phần tử/cá thể nào với lớp B. Phát biểu Disjoint Classes(A B C) có nghĩa là mỗi lớp trong đó disjoint với từng lớp còn lại (mutually disjoint). Phát biểu ABox dạng DisjointClasses(Vegetarian Carnivore) là sai vì Dora vừa thuộc Vegetarian vừa thuộc Carnivore dựa vào phát biểu Individual: Dora Types: Vegetarian, Carnivore.

4.2.3 Các phát biểu ABox xung đột với nhau

- Trường hợp này thì tương tự như nguyên nhân ở trên nhưng khác ở chỗ là lần này sự mâu thuẫn nằm trong các biểu ở cấp độ cá thể (ABox).
- Ví dụ:

Individual: Dora

Types: Vegetarian, not Vegetarian

- Dễ thấy được sự mâu thuẫn trong trong phát biểu trên vừa yêu cầu Dora là Vegetarian vừa yêu cầu nó không phải Vegetarian.

4.2.4 Phát biểu xung đột với nghĩa "oneOf" (All TBox)

- Phát biểu bao gồm hoặc một trong (oneOf trong cú pháp của OWL) cho phép sử dụng các cá thể trong khai báo phát biểu ABox, sự kết hợp này có thể dẫn đến sự thiếu nhất quán.

- Lấy ví dụ sau:

```
Class: MyFavouriteCow
    EquivalentTo: {Dora}
Class: AllMyCows
    EquivalentTo: {Dora, Daisy, Patty}
DisjointClasses: MyFavouriteCow, AllMyCows
```

- Phần đầu tiên của các phát biểu trên tất cả các thể thuộc lớp **MyFavouriteCow** phải tương đương với cá thể tên Dora, nói cách khác là **SameIndividual** với Dora. Phần thứ hai cũng tương tự tất cả các cá thể thuộc lớp **AllMyCows** buộc phải tương đương với một trong 3 cá thể tên Dora, Daisy hoặc Patty. Do 2 phát biểu trên chúng ta đã nói Dora thuộc cả 2 lớp **MyFavoriteCow** và **AllMyCows** nên mâu thuẫn với phát biểu cuối cùng khi nói 2 lớp này không có chung một cá thể nào. Vì vậy dẫn tới ontology bị thiếu nhất quán (inconsistent).

4.2.5 Không có khả năng khởi tạo bất kì cá thể nào (all TBox)

- Ví dụ:

```
Vegetarian or not Vegetarian
SubClassOf: Cow and not Cow
```

- Đây chỉ là một ví dụ đơn giản để minh họa cho trường hợp này. Thực tế sẽ ít người dùng nào tạo ra một phát biểu ngớ ngẩn như vậy nhưng nó vẫn có khả năng xảy ra khi phát biểu trên là kết quả từ suy luận (reasoning) của những phát biểu lớn và phức tạp hơn.
- Có thể giải thích ví dụ trên như sau. Đầu tiên để đáp ứng ý nghĩa dòng đầu tiên yêu cầu cá thể vừa là **Vegetarian** hoặc không phải **Vegetarian** - bất kỳ phát biểu nào dạng này, "cá thể thuộc hoặc không thuộc một lớp" chính là tất cả cá thể xuất hiện trong ontology. Dòng thứ hai yêu cầu cá thể vừa là Cow vừa không phải là Cow, phát biểu này rơi vào một trong các nguyên nhân vừa nêu ở trên. Tổng hợp lại chúng yêu cầu tất cả cá thể vừa là Cow vừa không phải Cow, điều này gây ra mâu thuẫn trên toàn ontology do phát biểu đầu tiên chỉ tới tất cả các cá thể.

Kết luận Trên đây chúng em đã liệt kê những nguyên nhân phổ biến dẫn đến thiếu nhất quán qua những ví dụ đã được đơn giản hoá để dễ dàng nắm bắt được đâu là căn nguyên gây ra sự mâu thuẫn về logic. Trên thực tế với những ontology có số lượng phát biểu lớn và phức tạp rất khó để người dùng có thể nhận diện được đâu là nguyên nhân

chính xác gây ra mâu thuẫn, do vậy sự ra đời của một công cụ giúp chúng ta phát hiện chính xác nguyên nhân gây lỗi là rất cần thiết. Vì vậy trong nội dung chương 2, chúng em sẽ đề cập tới Ontology Debugging một khía cạnh rất được chú trọng khi số lượng phát biểu của ontology ngày càng tăng.

Chương 5

Giải pháp để sửa chữa inconsistent ontology

- Như đã được đề cập trong chương 1, trong các nguyên nhân dẫn đến tính thiếu nhất quán (*Inconsistency*) trong ontology thì **Unsatisfiable Class** (lớp không thỏa về tính logic) là nguyên nhân nếu có thể được phát hiện sớm để loại bỏ hoặc sửa lại các phát biểu gây mâu thuẫn thì giúp cho ontology tránh bị inconsistent.
- Đã có rất nhiều nghiên cứu thành công trong việc tìm và phát hiện lỗi (các phát biểu mâu thuẫn) trong ontology. Trong đó có một nghiên cứu nổi bật[?], không chỉ có khả năng phát hiện gần như chính xác các nguyên nhân gây lỗi mà còn được đưa ra các giải pháp tối ưu* để sửa lỗi. Nghiên cứu này đã được ứng dụng để đưa ra các giải thích về các lớp không thỏa về nghĩa (*unsatisfiable classes*) trong bộ thực viện lập trình ontology thông dụng hiện nay là OWL-API[?]. Sau đây chúng em xin được trình bày lại những điểm quan trọng trong nghiên cứu vừa được đề cập**

* Tối ưu có nghĩa là hạn chế tối đa các thay đổi về ý nghĩa mà việc xóa hoặc thay đổi phát biểu mâu thuẫn có thể gây ra cho các phát biểu khác (*other axioms*) trong ontology.

** Mọi quan điểm và ý tưởng trình bày ở phần sau của chương này đều thuộc của các tác giả bài báo [?] và [?]. Chúng em chỉ trình bày lại sau khi đã đọc và nắm được ý tưởng chính yếu của bài báo.

5.1 Mục tiêu của việc debugging ontology

Mục tiêu chính của việc debugging ontology gồm hai phần quan trọng. Thứ nhất, với một ontology có số lượng lớn các lớp unsatisfiable, cần tìm và nhận dạng được nguyên nhân gây ra mâu thuẫn và các lớp bị ảnh hưởng bởi sự mâu thuẫn đó trong ontology. Thứ hai, cho biết trước một Unsatisfiable Class cụ thể, trích xuất và trình bày cho người sử dụng ontology (*modeler*) một tập hợp tối thiểu các phát biểu (*minimal set of axioms*) từ ontology hay nguyên nhân chính xác chịu trách nhiệm trong việc gây ra sự mâu thuẫn về logic.

5.2 Khái niệm và các kỹ thuật cần biết

Các hệ thống Description Logic thường cung cấp một tập hợp các tác vụ suy luận đã được chuẩn hóa như phân loại các khái niệm (*concept classification*), kiểm tra tính đáp ứng về logic (*concept satisfiability*) và kiểm tra tính nhất quán của knowledge base (KB). Hầu hết các reasoner thông dụng hiện nay đều buộc phải cung cấp đủ 3 tác vụ nêu trên, nhưng tất cả chúng đều không thân thiện với người dùng. Do tất cả những gì chúng ta biết được đều là kết quả (hay output) từ sự suy luận (reasoning) của reasoner.

Để giúp cho các tác vụ suy luận (reasoning) trở nên thân thiện với người dùng hơn, một hệ thống DL-based Knowledge Representation (KR) phải mở rộng thêm các lựa chọn về các tác vụ không nằm trong tiêu chuẩn của DL. Một ví dụ cụ thể là việc tạo ra các giải thích tại sao một lớp lại bị reasoner đánh giá là unsatisfiable. Thêm một tình huống mà người dùng cần được giải thích là tại sao reasoner đánh giá một lớp là lớp con của một lớp khác - đâu là lý do. Việc ra đời tác vụ giải thích nguyên nhân và kết quả là thật sự cần thiết trong bối cảnh sự phát triển nhanh của Semantic Web và cộng đồng người dùng/nhà phát triển Ontology ngày càng tăng nhanh.

5.2.1 Dịch vụ Axiom Pinpointing

Axiom Pinpointing service chính là dịch vụ có khả năng thực hiện tác vụ giải thích vừa được đề cập, với một KB và bất kì kết quả suy luận nào từ KB, dịch vụ này sẽ trả về tập các chứng minh/giải thích cho suy luận đó bằng những phát biểu đã được khai báo trong KB.

Có thể giải thích ngắn gọn như sau [?, p. 2], cho một phát biểu kết quả họ SHOIN α được suy ra từ một knowledge base K , một kiểm chứng (justification) cho α trong K là một phần tối thiểu $K' \subseteq K$ chịu trách nhiệm cho α xảy ra. Kiểm chứng K' là tối thiểu

với điều kiện α là một kết quả logic được suy ra từ K' , hay nói cách khác K' tối tiểu khi và chỉ khi bất kì tập con nào của K' đều không suy ra được α . Nói chung có thể tồn tại nhiều giải thích/chứng minh cho α trong K .

Sau đây là một ví dụ cho ý tưởng vừa nêu. Cho KB K với các phát biểu như sau:

1. $A \subseteq B \cap C$
2. $B \subseteq \neg E$
3. $A \subseteq D \cap \exists R.E$
4. $D \subseteq C \cap \forall R.B$

Trong KB trên, A, B, C, D, E là atomic concepts và R là atomic role. Chúng ta sẽ dùng số thứ tự của từng câu phát biểu trên thay vì lặp lại nguyên văn.

Từ các phát biểu trên ta có $K \models (A \subseteq C)$. Tuy nhiên, điều kiện cần và đủ để suy ra được một kết quả tương tự từ 2 phần nhỏ hơn của KB K là $K_1 = 1$ và $K_2 = 3, 4$. Chúng ta nói K_1 và K_2 là các kiểm chứng cho kết luận nói C là tập con của A - $A \subseteq C$.

KB trong ví dụ vừa nêu được xem là khá nhỏ, qua đó dễ dàng nhận ra lợi ích đáng kể khi số lượng phát biểu trong KB tăng lên vài trăm hay vài ngàn phát biểu. Bằng cách nhận dạng chính xác các tập tối tiểu chứa các phát biểu khẳng định (asserted) là những giả thiết cho kết quả được suy ra, dịch vụ này có thể được dùng để cô lập, đánh dấu và giải thích nguyên nhân hoặc cơ sở của các kết quả suy luận. Điều này cực kì quan trọng trên khía cạnh debugging, lấy ví dụ trường hợp cần giải thích là một Unsatisfiable Class/Concept, dịch vụ này sẽ khám phá tất cả và chỉ những phát biểu là nguyên nhân gây lỗi. Trong trường hợp vừa nêu, tìm kiếm tất cả các kiểm chứng là rất cần thiết vì để sửa lại unsatisfiable class cần loại bỏ ít nhất một phát biểu trong tập các phát biểu tối tiểu nguyên nhân gây lỗi MUPS - sẽ được đề cập trong mục bên dưới.

Tuy nhiên, dịch vụ axiom pinpointing chúng ta đề cập có một giới hạn là nó chỉ làm việc ở mức độ giữa các phát biểu với nhau, chúng vẫn chưa phân biệt được phần cụ thể nào của phát biểu mới là nguyên nhân cần và đủ để giải thích cho kết quả suy luận. Lấy lại ví dụ vừa nêu trên KB K , lớp B trong giao của $B \cap C$ trong phát biểu 1, không phải là giả thiết cần để suy ra $A \subseteq C$. Tương tự, $\exists R.E$ và $\forall R.B$ trong phát biểu 3 và 4 không phải điều kiện cần để suy ra được $A \subseteq C$.

Do vậy, việc quan tâm xem phần nào của phát biểu mới chính là giả thiết/nguyên nhân của kết quả suy luận rất quan trọng trong nhiều trường hợp, đặc biệt khi sửa chữa một phát biểu gây lỗi thì việc sửa lại một phần của phát biểu sẽ hạn chế sự mất mát về ý nghĩa của ontology hơn là xóa nó đi.

Để đáp ứng yêu cầu này, họ đề định nghĩa một *hàm chia nhỏ KB*. Ý tưởng là viết lại một phát biểu bất kì trong KB thành những dạng tập gồm các phát biểu nhỏ và

đơn giản hơn với ý nghĩa tương đương. Sau đó, sử dụng *Axiom Pinpointing Service* lên những tập những phát biểu trong KB K_s đã được viết lại từ K để tìm kiếm nguyên nhân hay giải thích cho kết quả suy luận.

Lấy phát biểu 1 trong ví dụ trên:

$$A \subseteq B \cap C \text{ (1) được viết lại thành } A \subseteq B, A \subseteq C \text{ (1*)}$$

Dễ dàng thấy phần $A \subseteq C$ trong 1* chính là điều phải chứng minh cho $K \models (A \subseteq C)$, những phần còn lại không cần thiết. Tương tự, ta viết lại (3) và (4) như sau:

$$\begin{aligned} A \subseteq D \cap \exists R.E &\Leftrightarrow A \subseteq D, A \subseteq \exists R.E \\ D \subseteq C \cap \forall R.B &\Leftrightarrow D \subseteq C, D \subseteq \forall R.B \end{aligned}$$

Bây giờ, điều kiện cần và đủ để chứng minh $K \models (A \subseteq C)$ là $A \subseteq D$ và $D \subseteq C$. Tuy nhiên, trong một vài trường hợp "hàm chia nhỏ KB" này đòi hỏi phải giới thiệu ra một tên lớp mới, viết giới thiệu tên lớp mới này chỉ phục vụ cho mục đích viết lại phát biểu. Ví dụ:

$$A \subseteq \exists R. (C \cap D) \text{ không tương đương với } A \subseteq \exists R.C, A \subseteq \exists R.D$$

Để chia nhỏ phát biểu trên chúng ta sẽ giới thiệu một tên lớp mới, gọi là E . Như vậy ta có:

$$A \subseteq \exists R. (C \cap D) \Leftrightarrow A \subseteq \exists R.E, E \subseteq C, E \subseteq D, C \cap D \subseteq E$$

Để thực hiện được cái gọi là "hàm chia nhỏ KB" các bài báo [?] và [?] đã đề xuất các giải thuật với tiêu chí xác định các phát biểu chứng minh một cách đầy đủ và chính xác. Các giải thuật này có thể được chia thành 2 nhóm:

1. *Reasoner Dependent(or Glass-box) Algorithm* Đây là nhóm các giải thuật xây dựng trên quy trình đưa ra quyết định Tableau dành cho Description Logic. Tuy nhiên, để áp dụng các giải thuật loại này trong thực tế đòi hỏi phải có những chỉnh sửa đáng kể bên trong quy trình suy luận những DL reasoner hiện nay.
2. *Reasoner Independent(or Black-box) Algorithm* Nhóm này chỉ sử dụng các DL reasoner cho những tác vụ kiểm tra lại kết quả suy luận khi đã viết lại KB K thành K' , chúng không đòi hỏi phải chỉnh sửa lại các cách hoạt động của reasoner. Reasoner lúc này có chức năng như một "chiếc hộp đen" chấp nhận các input là lớp/các phát biểu đã được viết lại hoặc một KB K' viết lại từ KB K , sau đó trả

về output là một câu trả lời xác nhận hay phủ định rằng các lớp và các phát biểu này có là tập tối tiểu để chứng minh cho kết quả suy luận hay không. Ví dụ trong trường hợp:

$$A \subseteq B \cap C \Leftrightarrow A \subseteq B, A \subseteq C$$

Các inputs của reasoner sẽ lần lượt sau mỗi vòng là

- (a) $A \subseteq B, A \subseteq C$
- (b) $A \subseteq B$
- (c) $A \subseteq C$

Giải thuật sẽ lần lượt loại bỏ từng phát biểu một (sau mỗi vòng) để xem những phát biểu còn lại có đủ chứng minh $K \models (A \subseteq C)$. Đến khi nào giải thuật không tồn tại tập phát biểu nào đủ để chứng minh $K \models (A \subseteq C)$ thì sẽ dừng vòng lặp.

Kết luận: trên đây chỉ là những bước hoạt động cơ bản nhất của Axiom Pinpointing Service và Blackbox Algorithm xin đọc thêm [?]. Các giải thuật và dịch vụ này cũng đã được áp dụng trong package com.clarkparsia.owlapi.explanation của [?]

5.2.2 Minimal Unsatisfiability Preserving Sub-TBoxes (MUPS)

Khái niệm MUPS lần đầu được giới thiệu trong[?].Như đã được đề cập trong phần đầu của mục này, một MUPS thật ra chính là một phần nhỏ nhất của KB K mà trong đó lý giải tại sao một lớp lại unsatisfiable, nói cách khác một MUPS là một tập tối tiểu các phát biểu mà trong đó các phát biểu này giải thích chính xác nguyên nhân gây ra mâu thuẫn về logic(unsatisfiable). Một lớp unsatisfiable có thể có nhiều MUPS trong KB K (hay cụ thể là trong ontologies). Ví dụ có KB K_α với những phát biểu như sau:

1. $S \equiv A \cap \exists R.B$
2. $S \subseteq \exists R.(C \cap D)$
3. $(C \cap D) = \emptyset$

Dựa vào các phát biểu trên ta thấy S unsatisfiable. MUPS của S từ K_α là:

$$S \subseteq \exists R.(C \cap D) \quad (2)$$

$$(C \cap D) = \emptyset \quad (3)$$

Để sửa lại một lớp không đáp ứng (*unsatisfiable class*) chúng ta cần loại bỏ tối thiểu ít nhất một phát biểu từ từng tập các phát biểu tối thiểu MUPS lý giải cho *unsatisfiable class* đó. Trong ví dụ vừa rồi do chỉ có 1 MUPS, ta bỏ tất cả các phát biểu trong MUPS xuất hiện trong KB K_α thì S sẽ lại *satisfiable*.

5.3 Các bước sửa chữa các phát biểu bị lỗi

5.3.1 Tìm tất cả các MUPS của một Unsatisfiable Class

Như vừa nói ở trên MUPS thật ra chính là một phần nhỏ nhất trong KB khiến cho một lớp *unsatisfiable*. Do vậy tìm và xác định MUPS chính là tìm và xác định các tập tối thiểu các phát biểu cho một lớp được suy luận là *unsatisfiable*. Chúng ta sẽ sử dụng *Axiom Pinpointing Service*[?] để tìm MUPS với các bước tương tự đã được mô tả chi tiết trong mục trên. Nhiệm vụ tìm kiếm *precise* MUPS của lớp không đáp ứng trong KB K được đơn giản hóa thành vấn đề tìm MUPS trong những phiên bản đã được tách nhỏ trong KB K_s .

5.3.2 Chiến thuật xếp hạng các phát biểu (*Axioms*)

Đây là một giai đoạn khá quan trọng trong quá trình chỉnh sửa lại các phát biểu gây lỗi, quyết định xem nên loại bỏ phát biểu nào từ các MUPS để lớp/khái niệm được *satisfiable*.

Với mục tiêu này, một nhân tố đáng quan tâm là các phát biểu trong MUPS có thể được *xếp hạng* dựa theo mức độ quan trọng của chúng. Việc sửa chữa các nguyên nhân gây lỗi được trở thành một vấn đề cần được tối ưu để đáp ứng các tiêu chí vừa phải loại bỏ tất cả các lỗi gây ra tính thiếu nhất quán trong ontology, trong khi vẫn chắc chắn rằng những phát biểu có thứ hạng cao, nói cách khác là có giá trị quan trọng về nghĩa sẽ được ưu tiên giữ lại và các phát biểu có thứ hạng thấp nhất sẽ bị loại bỏ.

Tiêu chí đơn giản nhất để xếp hạng các phát biểu là đếm số lần chúng xuất hiện trong MUPS từ những lớp *unsatisfiable* xuất hiện trong một ontology. Nếu một phát biểu xuất hiện trong n MUPS khác nhau (trong từng tập phát biểu của MUPS), bỏ đi phát biểu đó sẽ đảm bảo rằng n lớp/khái niệm được *satisfiable*. Số lần phát biểu xuất hiện càng nhiều, thứ hạng của nó càng thấp.

Ngoài tần suất xuất hiện của phát biểu trong MUPS, chúng ta cũng có thể quan tâm đến những yếu tố sau để đưa vào tiêu chí xếp hạng:

- Tác động lên ontology khi loại bỏ phát biểu hoặc thay đổi nội dung phát biểu - cần phải nhận diện được những tác động tối thiểu (*minimal impact*) gây ra thay đổi.
- Tự xây dựng những test cases cụ thể để xếp hạng các phát biểu dựa theo tiêu chí của người dùng tự đề ra.
- Dựa trên những metadata của phát biểu như tác giả, độ tin cậy của nguồn tài liệu, timestamp, etc.
- Sự liên quan tới ontology ở khía cạnh phát biểu được sử dụng vào mục đích gì và sử dụng như thế nào.

Lưu ý: Chi tiết về cách áp dụng từng tiêu chí xếp hạng trên xin đọc [?].

5.3.3 Tạo ra các giải pháp sửa lỗi

Qua các phần trên, chúng ta đã biết được làm thế nào để tìm MUPS cho một lớp unsatisfiable bằng Axiom Pinpointing Service trong một OWL-DL ontology và thấy được một loạt các tiêu chí để xếp hạng phát biểu trong MUPS. Bước tiếp theo là tạo ra một kế hoạch sửa lỗi (hay một loạt các thay đổi trong ontology) để sửa các lỗi trong một tập các lớp/khái niệm bị unsatisfiable, với các dữ kiện đã có qua các bước trên như các MUPS tìm được và thứ hạng các phátrepair

Điều chỉnh giải thuật Reiter Giải thuật Hitting Set của Reiter[?], đưa ra nhằm để xác định căn nguyên(*root cause*) của một vấn đề từ một bộ(*collection*) gồm nhiều tập hợp đựng độ chứa các nguyên nhân dẫn tới vấn đề, giải thuật này sẽ tạo ra những tập tối thiểu (*minimal hitting set*) chứa các nguyên nhân gây ra vấn đề. Một tập hợp đựng độ (*hitting set*) trong một bộ **C** các tập hợp là tập hợp giao (có chung phần tử) với từng tập hợp trong **C**. Một tập hợp đựng độ là tối thiểu nếu không có bất kì tập con nào của nó lại là một tập đựng độ cho **C**. Trong trường hợp của chúng ta, bộ **C** chứa các HST chính là các MUPS tìm được trong ontology.

Ý tưởng là áp dụng giải thuật Reiter để tìm ra tập tối thiểu các phát biểu gây lỗi từ các MUPS đã tìm được, rồi loại bỏ tất cả các phát biểu trong tập đựng độ tối thiểu từ đó giúp loại bỏ từng phát biểu gây lỗi xuất hiện trong từng tập phát biểu từng MUPS và cuối cùng giúp cho sửa chữa được cho lớp/khái niệm được satisfiable. Nguyên lý tương tự cũng được áp dụng cho việc giải pháp sửa lỗi ngoại trừ cần phải điều chỉnh lại giải thuật HS để nó có thể hoạt động dựa trên thứ hạng của các phát biểu.

Cho một bộ **C** gồm những tập đựng độ, giải thuật Reiter giới thiệu một khái niệm về hitting set tree (HST), là một cấu trúc cây có số cạnh nhỏ nhất và số node nhỏ nhất,

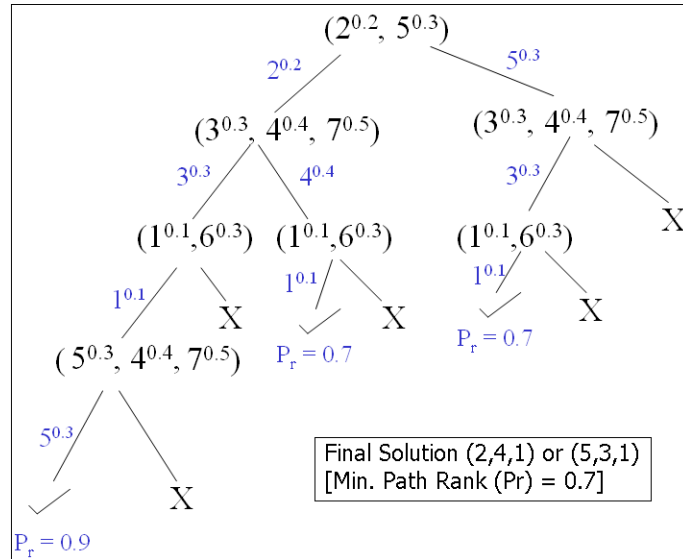
với cạnh và node đều được dán nhãn (labeled). Một node n trong HST được dán nhãn bởi dấu tick (✓) nếu C rỗng, ngược lại node này sẽ được dán nhãn bởi bất kì tập hợp $s \in C$. Với mỗi node n , ta có $H(n)$ là tập gồm các nhãn của cạnh (edge labels) trên đường đi từ gốc cây tới n (root to n); và nhãn cho n là bất kì tập $s \in C$, thỏa điều kiện $s \cap H(n) \leftarrow \emptyset$, nếu có một tập hợp nào như vậy tồn tại. Nếu n được dán nhãn bởi một tập s , thì với từng $\sigma \in s$, n có một node kế cận là n_σ nối với n bởi một cạnh được dán nhãn bằng σ . Với bất kì node nào được dán nhãn bằng ✓, tập chứa các nhãn mô tả đường đi (theo cạnh) của node này tới gốc cây là một tập đựng độ(hitting set) của C . Khi tạo ra HST từ gốc, nếu trong quá trình tìm kiếm phát hiện được giải một giải pháp tối ưu hiện thời, thì quá trình sẽ được kết thúc sớm hơn, đánh dấu bằng một bằng dấu chéo (✗) trên nhãn của node.

Áp dụng vào trường hợp của chúng ta, MUPS của các lớp unsatisfiable tương đương với các tập hợp đựng độ. Tuy nhiên, trong giải thuật HST bình thường được tối ưu theo tiêu chí đường đi ngắn nhất, thay vì đường đi ngắn nhất chúng ta sẽ sử dụng thứ hạng nhỏ nhất (minimal path rank), nói cách khác tổng thứ hạng của các phát biểu trong $H(n)$ sẽ phải nhỏ nhất. Thêm nữa, là trong giải thuật HST cơ bản, không tồn tại khái niệm lựa chọn một phát biểu trong những phát biểu khác trong khi xây dựng cạnh của HST, trong khi chúng ta có thể sử dụng thứ hạng của các phát biểu trong lúc quyết định lựa chọn để thu hẹp không gian tìm kiếm, hay nói dễ hiểu là trong mỗi giai đoạn xây cạnh chúng ta sẽ chọn phát biểu có thứ hạng thấp nhất.

Hình 2.1 Thể hiện một HST của một collection C chứa các phát biểu từ 1 - 7 $C = 2,5, 3,4,7, 1,6, 4, 5, 7, 1, 2, 3$ với thứ hạng của các phát biểu từ 1 - 7 như sau: $r(1) = 0.1, r(2) = 0.2, r(3) = 0.3, r(4) = 0.4, r(5) = 0.3, r(6) = 0.3, r(7) = 0.5$, trong đó $r(x)$ là hạng của phát biểu x . Thứ hạng này được tính ra dựa trên những yếu tố được đề cập ở phần 2.3.2 như tần suất xuất hiện, tác động ngữ nghĩa, etc. mỗi tiêu chí được đánh giá riêng biệt, nếu cần chúng ta có thể quy ước một hệ số để đánh giá tất cả cùng một lúc. Số mũ trên từng phát biểu chính biểu diễn hạng của phát biểu đó, và P_r là *path rank* được tính bằng tổng hạng của các phát biểu nằm trên đường đi (theo cạnh) từ gốc tới một node. Ví dụ, cạnh cận trái nhất có *path rank*: $P_r = 0.2 + 0.3 + 0.1 + 0.3 = 0.9$.

Như được thể hiện trong hình, bằng cách chọn phát biểu có hạng thấp nhất trong từng tập trong khi xây cạnh của HST, giải thuật chỉ tạo ra 3 hitting sets, 2 trong số đó tối tiểu, trong khi hạn chế được một số lượng lớn số lần kiểm tra đường đi, (thể hiện bằng ✗). Giải pháp sửa lỗi được tìm ra trong tập có P_r nhỏ nhất là 2,4,1 hoặc 5,3,1.

Tuy vậy, có một hạn chế khi sử dụng quy trình vừa nêu trên để tạo ra kế hoạch sửa lỗi, như phân tích tác động ngữ nghĩa của phát biểu chỉ được thực hiện ở cấp độ là một phát biểu đơn lẻ, trong khi một loạt tác động khác chưa được tính tới mỗi lần một HS được tìm thấy. Điều này có thể dẫn tới một giải pháp kém tối ưu. Ví dụ:



HÌNH 5.1: Giải thuật HST được chỉnh sửa dựa theo thứ hạng của phát biểu

1. DisjointClasses(Car Plane Ship) EquivalentClass(FlyingCar (Car and Plane))
- 2.

Trong ví dụ trên, bỏ Plane ra khỏi phát biểu (1) sẽ hạn chế mất mát về nghĩa hơn là xóa hết cả phát biểu (1), vì có thể disjoint giữa Car và Ship có thể được sử dụng đâu đó trong ontology mà chưa được tính đến.

Để khắc phục hạn chế này, một chỉnh sửa khác được đưa ra là cứ mỗi lần tìm ra hitting-set(HS), chúng ta sẽ tính lại thứ hạng của đường đi (path-rank) cho HS dựa trên một loạt tác động của các phát biểu trong hitting-set. Giải thuật bây giờ sẽ tìm được giải pháp tối thiểu được path-ranks mới.

Trên đây là ý tưởng cơ bản của giải thuật HST, ngoài ra còn những mục về cải thiện các giải pháp sửa lỗi và gợi ý các phát biểu sửa lỗi xin đọc thêm ở [?].

5.4 Ứng dụng HST để xác định tất cả kiểm chứng cho kết quả suy luận[?]

Ngoài ứng dụng vừa được đề cập ở trên, Hitting Set Tree còn được áp dụng để tìm tất cả các giải thích cho một kết quả suy luận, giống như một trong chức năng chính của Axiom Pinpointing Service được đề cập lúc này. Tuy nhiên, không giống như khái niệm HST của Reiter vừa được giới thiệu ở trên, ở đây chúng ta sẽ có HST với quy ước như sau.

Với ontology $\beta \models \eta$, một cây hitting set (HST) cho η trong β là một cây hữu hạn, bao gồm node được dán nhãn bằng các kiểm chứng(justifications) hay các phát biểu

chứng minh $\beta \models \eta$ và cạnh được đánh dấu với phát biểu trong β . Từng non-leaf(không phải lá) node v nối với một node kế cận v' qua một cạnh được dán nhãn với một phát biểu α với α nằm trong nhãn của v , nhưng không nằm trong nhãn của v' . Nhãn của v' có thể là một tập hợp rỗng, trong trường hợp đó v' phải là node lá (leaf node). Ngoài ra, với bất kì node v'' , tập chứa các phát biểu dán nhãn cho đường đi từ v'' tới gốc cây (tree root) không giao với kiểm chứng(hay các phát biểu chứng minh) dán nhãn v'' .

Quá trình xây dựng HST có thể được thực hiện bằng các giải thuật *breadth first* hay *depth first*. Dù được sử dụng theo cách nào, các nguyên lý và các luật để tạo và dán nhãn cạnh, node trên cây đều như nhau. Khi mở rộng cây từ node v đến một node mới v' quy trình cơ bản đều diễn ra như sau:

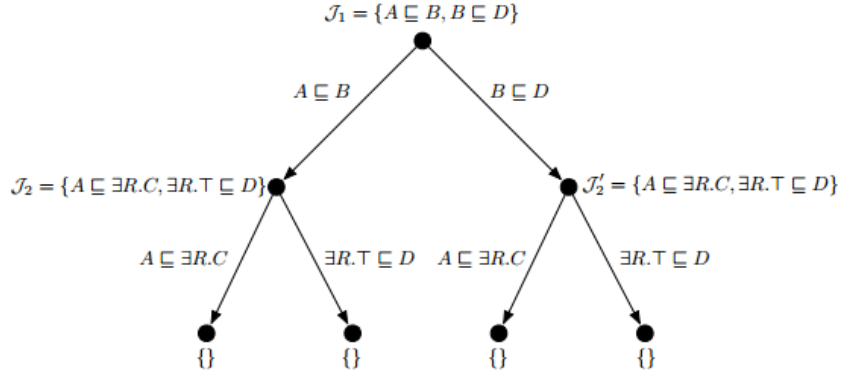
1. Chọn một phát biểu α nằm trong nhãn của v nhưng không dán nhãn một cạnh nối v tới bất kì node kế cận nào.
2. Gọi S là hội của (union of) α và tập những phát biểu nằm trên cạnh, tạo thành đường đi từ v tới root node. Loại bỏ S khỏi β ta được β' .
3. Nếu η thỏa $\beta' \models \eta$ thì tìm một kiểm chứng (justification) J cho η trong β' . Nếu $\beta' \not\models \eta$ thì gán $J = \emptyset$.
4. Tạo một node mới v' và dán nhãn cho v' bằng tập J ở bước trên.
5. Tiếp tục mở rộng HST theo một hướng bằng cạnh $e = (v, v')$ tương tự như bước 1 cho tới khi không tìm được tập $J = \emptyset$ như ở bước 2.
6. Đưa các phát biểu trong S trở lại β .

Ví dụ sau mô tả lại quy trình trên, cho ontology β với các phát biểu sau:

1. $A \subseteq B$
2. $B \subseteq D$
3. $A \subseteq \exists R.C$
4. $\exists R.\top \subseteq D$

Trong đó $\eta = A \subseteq D$. HST cho $\beta \models A \subseteq D$ được thể hiện ở hình 2.2. Bắt đầu di chuyển tại root node, node được dán nhãn bởi J_1^* , mở rộng HST về phía bên trái bằng cách chọn phát biểu $A \subseteq B$ trong J_1 với điều kiện $A \subseteq B$ chưa dán nhãn bất kì cạnh nào nối với root node sau đó loại bỏ phát biểu $A \subseteq B$ khỏi β và tính toán lại kiểm chứng (justifications) thỏa $\langle \beta \setminus \{A \subseteq B\} \rangle \models A \subseteq D$. Trong trường hợp này, chúng ta tìm được kiểm chứng J_2 trong $\beta \setminus \{A \subseteq B\}$, giải thích được tại sao $A \subseteq D$. Tiếp tục đi

về phía bên trái của node J_2 , chọn $A \subseteq \exists R.C$ trong J_2 tương tự cách chọn ở bước 1. Sau đó tìm kiếm các kiểm chứng trong $\beta' \equiv \beta \setminus \{A \subseteq \exists R.C, A \subseteq B\}$, kết quả là chúng ta không tìm được kiểm chứng trong β' giải thích cho $A \subseteq D$ hay có thể nói là $\beta' \not\models (A \subseteq D)$, do vậy node kế được dán nhãn bằng \emptyset vì lúc này $J = \emptyset$ (bước 3). Mỗi lần như vậy ta tìm ra leaf-node, chúng ta sẽ đưa S trở lại β và bắt đầu lại quá trình tìm kiếm.



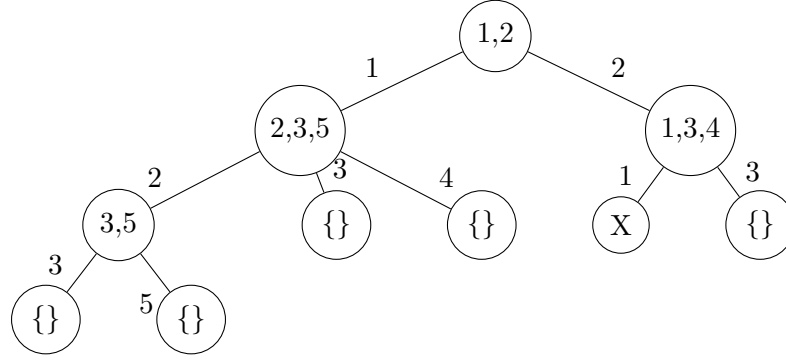
HÌNH 5.2: Hitting Set Tree dùng để tìm kiếm kiểm chứng

Quá trình lặp lại cho đến khi chúng ta không còn tạo được node mới nào, tại thời điểm này thì việc xây dựng HST cũng hoàn tất. *Tất cả* các kiểm chứng để giải thích cho $\beta \models \eta$ chính là nhãn của các node trong HST. Thêm một điều nữa, là tất cả các đặc điểm đơn giản nhất để chứng minh $\beta \models \eta$ nằm trên các cạnh từ leaf-node tới root-node của cây. Ví dụ vừa nêu trên chỉ biểu diễn những bước cơ bản nhất để xây dựng một HST nhưng quan tâm đến bất kì khả năng tối ưu hóa nào cho giải thuật. Để đạt được một hiệu năng chấp nhận được khi ứng dụng trong thực tế thì 2 giải pháp tối ưu sẽ được nêu ra sau đây.

Early Path Termination - Trong phiên bản chưa tối ưu ở trên, một node n bất kì khi tạo cạnh mới với phát biểu thuộc tập $H(n)$, với $H(n)$ là tập phát biểu dán nhãn n , phát biểu trên cạnh mới này không được nằm trên bất kì cạnh nào nối n với một node kế cận (successor nodes). Chúng ta gọi các nodes có khả năng mở rộng (tạo được cạnh mới) là *open nodes*, ngược lại các leaf-nodes không có khả năng mở rộng là *closed nodes*. Để thực hiện tối ưu hóa, sẽ có trường hợp mà những nodes không phải leaf-nodes có thể được dán nhãn bởi những tập khác \emptyset nhưng vẫn sẽ được đánh dấu là *closed nodes*. Trong tình huống này, đường đi từ *closed node* tới root được chỉ định là *early termination* - kết thúc sớm. Để phát hiện được *early termination* chúng ta sẽ làm theo quy trình sau: Nếu một HST T chứa một open node v_1 , có đường đi P_1 tới root node, có thêm một open node v_2 cũng trong T , có đường tới root node là P_2 . Nếu tập dán nhãn cho P_1 bằng với tập dán nhãn cho P_2 thì v_2 sẽ được đánh dấu là *closed node* và không cần thiết phải mở

* J_1, J_2, J_2' được tính ra nhờ giải thuật Black-box hoặc Glass-box được đề cập ở trên.

rộng thêm nữa. Ví dụ chúng ta có ontology O chứa các phát biểu sau $O = \{1, 2, 3, 4, 5\}$ và $O \models \alpha$ Chúng ta bắt đầu di chuyển root node với tập $\{1, 2\}$ là các phát biểu đầu tiên



HÌNH 5.3: Early Termination trong HST Explanation

tìm được trong O chứng minh được $O \models \alpha$, thực hiện tương tự các bước đã được miêu tả ở trên ta thu được node $\{3, 5\}$ là các phát biểu giải thích cho $O \models \alpha$, tới lúc này ta có thể thấy tập chữ đường đi theo cạnh từ node $\{3, 5\}$ tới root node là $P_1 = \{2, 1\}$. Nhìn về phía bên phải ta phát hiện khi mở rộng cạnh từ node $\{1, 3, 4\}$ ta thu được đường đi về root node là $P_2 = \{1, 2\}$, ta thấy $P_1 \equiv P_2$ do vậy nên khi dán nhãn cho node kế tiếp (được đánh dấu **X** cho closed node) chúng ta sẽ bỏ $\{1, 2\}$ khỏi O để được $O' = \{3, 4, 5\}$, sẽ có một node giống y như node $\{3, 5\}$ sẽ xuất hiện lần nữa ở node kế tiếp này nên việc kết thúc ở đây là cần thiết vì chúng ta sẽ tiếp kiểm được việc kiểm tra lại $\{3, 5\}$ như ở bên trái.

Justification Reuse - Cách quan trọng thứ hai để tối ưu là sử dụng lại các kiểm chứng. Trong phiên bản không tối ưu sử dụng ở ví dụ ontology β ở trên, kiểm chứng được tìm ra nhờ các giải thuật Blackbox hay Glassbox cho từng node v được thêm vào cây. Kiểm chứng hay tập các phát biểu được sử dụng để dán nhãn v , được tính toán dựa trên $O \setminus S$, với S là tập các nhãn trên đường đi từ v về root node. Thay vì dùng Glassbox hay Blackbox để tính J trong $O \setminus S$, chúng ta có thể làm theo cách sau: nếu HST chứa vài node khác v' mà được dán nhãn với kiểm chứng J , và S không giao (có phần tử chung) với J thì J có thể được sử dụng làm nhãn cho v . Lý do là vì khi $J \subseteq O$ và $S \cap J = \emptyset$ thì sẽ tồn tại trường hợp $J \subseteq O \setminus S$, từ đó J được tính như một kiểm chứng cho để có thể dán nhãn v . Sử dụng lại các phát biểu chứng minh (hay kiểm chứng) sẽ giúp tiết kiệm nhiều lời gọi hàm không cần thiết tới Blackbox hoặc Glassbox từ đó tăng được hiệu năng.

Kết luận Trong quá trình nghiên cứu về các nguyên nhân gây inconsistency trong ontology, chúng em đã tìm hiểu được nhiều giải pháp đã được nghiên cứu và ứng dụng. Chúng em cũng nắm được nguyên lý hoạt động những giải thuật và quy trình này như

HST Explanation ,Blackbox Algorithm [?]. Nguyên lý này được áp dụng trong các chứng năng giải thích của thư viện OWL-API [?].

Tài liệu tham khảo

- M. K. Bergman, “The open world assumption: Elephant in the room,” December 2009. [Online]. Available: <http://www.mkbergman.com/852/the-open-world-assumption-elephant-in-the-room/>
- “Ontology web language 2 overview,” 2012. [Online]. Available: <http://www.w3.org/TR/owl2-overview/>
- R. Stevens, “(i can’t get no) satisfiability.” [Online]. Available: <http://ontogenesis.knowledgeblog.org/1329>
- S. Bail, “Common reasons for ontology inconsistency.” [Online]. Available: <http://ontogenesis.knowledgeblog.org/1343>
- M. Horridge, “Justification based explanation in ontologies,” Ph.D. dissertation, The University of Manchester. [Online]. Available: <http://www.bcs.org/upload/pdf/dd-matthew-horridge.pdf>
- *Programming with the OWL API*, 2014. [Online]. Available: <https://github.com/owlcs/owlapi>
- *SWRL API*, 2014. [Online]. Available: <https://github.com/protegeproject/swrlapi/wiki>
- J. Squeda, “Introduction to: Open world assumption vs closed world assumption,” 2012. [Online]. Available: http://semanticweb.com/introduction-to-open-world-assumption-vs-closed-world-assumption_b33688
- P. F. Patel, Schneider, and I. Horrocks, “A comparison of two modelling paradigms in the semantic web,” *WWW2006*, May 2006. [Online]. Available: <http://www.cs.ox.ac.uk/people/ian.horrocks/Publications/download/2006/PaHo06a.pdf>
- “Introducing linked data and the semantic web.” [Online]. Available: <http://www.linkeddatatools.com/semantic-web-basics>
- “Semantic web standards by w3c.” [Online]. Available: http://www.w3.org/2001/sw/wiki/Main_Page

- [] “Semantic web.” [Online]. Available: http://en.wikipedia.org/wiki/Semantic_Web
- [] “Resource description framework.” [Online]. Available: http://www.w3.org/standards/techs/rdf#w3c_all
- [] “Rdf schema 1.1,” 2014. [Online]. Available: <http://www.w3.org/TR/2014/REC-rdf-schema-20140225/>
- [] “Mapping to rdf graph.” [Online]. Available: <http://www.w3.org/TR/2012/REC-owl2-mapping-to-rdf-20121211/>
- [] “Rdf/xml syntax specification.” [Online]. Available: <http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/>
- [] “Xml serialization.” [Online]. Available: <http://www.w3.org/TR/2012/REC-owl2-xml-serialization-20121211/>
- [] “Functional syntax specification.” [Online]. Available: <http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/>
- [] “Manchester syntax.” [Online]. Available: <http://www.w3.org/TR/2012/NOTE-owl2-manchester-syntax-20121211/>
- [] “Internationalized resource identifiers,” 2005. [Online]. Available: <http://www.ietf.org/rfc/rfc3987.txt>
- [] “Rdf concept.” [Online]. Available: <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>
- [] B. C.-G. A. Kalyanpur, B. Parsia and E. Sirin, “Repairing unsatisfiable concepts in owl ontologies,” 2007. [Online]. Available: <http://www.cs.ox.ac.uk/people/bernardo.cuencagrau/publications/repair.pdf>
- [] A. Kalyanpur, B. Parsia, B. Cuenca-Grau, and E. Sirin, “Axiom pinpointing: Finding (precise) justifications for arbitrary entailments in shoin (owl-dl),” UMIACS 2006, Tech. Rep., 2006.
- [] S. Schlobach and R. Cornet, “Non-standard reasoning services for the debugging of description logic terminologies,” IJCAI, 2003.
- [] R. Reiter, “A theory of diagnosis from first principles,” *Artificial Intelligence* 1987, 1987.