Khái niệm về ontology

Trong khoa học máy tính, một ontology là một mô hình dữ liệu biểu diễn một lĩnh vực và được sử dụng để suy luận về các đối tượng trong lĩnh vực đó và mối quan hệ giữa chúng. Ontology cung cấp một bộ từ vựng chung bao gồm các khái niệm, các thuộc tính quan trọng và các định nghĩa về các khái niệm và các thuộc tính này. Ngoài bộ từ vựng, ontology còn cung cấp các ràng buộc, đôi khi các ràng buộc này được coi như các giả định cơ sở về ý nghĩa mong muốn của bộ từ vựng, nó được sử dụng trong một miền mà có thể được giao tiếp giữa người và các hệ thống ứng dụng phân tán hỗn tạp khác.

Các ontology được sử dụng như là một biểu mẫu trình bày tri thức về thế giới hay một phần của nó. Các ontology thường miêu tả:

* Các cá thể: Các đối tượng cơ bản, nền tảng
* Các lớp: Các tập hợp, hay kiểu của các đối tượng
* Các thuộc tính: Thuộc tính, tính năng, đặc điểm, tính cách, hay các thông số mà các đối tượng có và có thể đem ra chia sẻ.
* Các mối liên hệ: Các con đường mà các đối tượng có thể liên hệ tới một đối tượng khác.

Bộ từ vựng ontology được xây dựng trên cơ sở tầng RDF và RDFS, cung cấp khả năng biểu diễn ngữ nghĩa mềm dẻo cho tài nguyên Web và có khả năng hỗ trợ lập luận.

Các phần tử trong ontolog

Các cá thể (Individuals) - Thể hiện

Các cá thể là các thành phần cơ bản, nền tảng của một ontology. Các cá thể trong một ontology có thể bao gồm các đối tượng cụ thể như con người, động vật, cái bàn… cũng như các cá thể trừu tượng như các thành viên hay các từ. Một ontology có thể không cần bất kỳ một cá thể nào, nhưng một trong những lý do chính của một ontology là để cung cấp một ngữ nghĩa của việc phân lớp các cá thể, mặc dù các cá thể này không thực sự là một phần của ontology.

Các lớp (Classes) - Khái niệm

Các lớp là các nhóm, tập hợp các đối tượng trừu tượng. Chúng có thể chứa các cá thể, các lớp khác, hay là sự phối hợp của cả hai.

Các ontology biến đổi tuỳ thuộc vào cấu trúc và nội dung của nó: Một lớp có thể chứa các lớp con, có thể là một lớp tổng quan (chứa tất cả mọi thứ), có thể là lớp chỉ chứa những cá thể riêng lẻ, Một lớp có thể xếp gộp vào hoặc bị xếp gộp vào bởi các lớp khác. Mối quan hệ xếp gộp này được sử dụng để tạo ra một cấu trúc có thứ bậc các lớp, thường là với một lớp thông dụng nhất kiểu Thing ở trên đỉnh và các lớp rất rõ ràng kiểu 2002, Ford ở phía dưới cùng.

Các thuộc tính (Properties)

Các đối tượng trong ontology có thể được mô tả thông qua việc khai báo các thuộc tính của chúng. Mỗi một thuộc tính đều có tên và giá trị của thuộc tính đó. Các thuộc tính được sử dụng để lưu trữ các thông tin mà đối tượng có thể có. Ví dụ, đối với một cá nhân có thể có các thuộc tính: Họ_tên, ngày_sinh, quê_quán, số_cmnd…

Giá trị của một thuộc tính có thể có các kiểu dữ liệu phức tạp.

Các mối quan hệ (Relation)

Một trong những ứng dụng quan trọng của việc sử dụng các thuộc tính là để mô tả mối liên hệ giữa các đối tượng trong ontology. Một mối quan hệ là một thuộc tính có giá trị là một đối tượng

nào đó trong ontology.
Một kiểu quan hệ quan trọng là kiểu quan hệ xếp gộp (subsumption). Kiểu quan hệ này mô tả các đối tượng nào là các thành viên của các lớp nào của các đối tượng.
Hiện tại, việc kết hợp các ontology là một tiến trình được làm phần lớn là thủ công, do vậy rất tốn thời gian và đắt đỏ. Việc sử dụng các ontology là cơ sở để cung cấp một định nghĩa thông dụng của các thuật ngữ cốt lõi có thể làm cho tiến trình này trở nên dễ quản lý hơn. Hiện đang có các nghiên cứu dựa trên các kỹ thuật sản sinh để nối kết các ontology, tuy nhiên lĩnh vực này mới chỉ hiện hữu về mặt lý thuyết.

Ngôn ngữ OWL
OWL (The Web Ontology Language) là một ngôn ngữ gần như XML dùng để mô tả các hệ cơ sở tri thức. OWL là một ngôn ngữ đánh dấu dùng để xuất bản và chia sẻ dữ liệu trên Internet thông qua những mô hình dữ liệu gọi là "ontology". Ontology mô tả một lĩnh vực (domain) và diễn tả những đối tượng trong lĩnh vực đó cùng những mối quan hệ giữa các đối tượng này. OWL là phần mở rộng về từ vựng của RDF và được kế thừa từ ngôn ngữ DAML+OIL Web ontology – một dự án được hỗ trợ bởi W3C. OWL biểu diễn ý nghĩa của các thuật ngữ trong các từ vựng và mối liên hệ giữa các thuật ngữ này để đảm bảo phù hợp với quá trình xử lý bởi các phần mềm. OWL được xem như là một kỹ thuật trọng yếu để cài đặt cho Semantic Web trong tương lai. OWL được thiết kế đặc biệt để cung cấp một cách thức thông dụng trong việc xử lý nội dung thông tin của Web. Ngôn ngữ này được kỳ vọng rằng sẽ cho phép các hệ thống máy tính có thể đọc được thay thế cho con người. Vì OWL được viết bởi XML, các thông tin OWL có thể dễ dàng trao đổi giữa các kiểu hệ thống máy tính khác nhau, sử dụng các hệ điều hành và các ngôn ngữ ứng dụng khác nhau. Mục đích chính của OWL là sẽ cung cấp các chuẩn để tạo ra một nền tảng để quản lý tài sản, tích hợp mức doanh nghiệp và để chia sẻ cũng như tái sử dụng dữ liệu trên Web. OWL được phát triển bởi nó có nhiều tiện lợi để biểu diễn ý nghĩa và ngữ nghĩa hơn so với XML, RDF và RDFS, và vì OWL ra đời sau các ngôn ngữ này, nó có khả năng biểu diễn các nội dung mà máy có thể biểu diễn được trên Web.

Các phiên bản của OWL
Hiện nay có ba loại OWL : OWL Lite, OWL DL (description logic), và OWL Full.
OWL Lite: hỗ trợ cho những người dùng chủ yếu cần sự phân lớp theo thứ bậc và các ràng buộc đơn giản. Ví dụ: Trong khi nó hỗ trợ các ràng buộc về tập hợp, nó chỉ cho phép tập hợp giá trị của 0 hay 1. Điều này cho phép cung cấp các công cụ hỗ trợ OWL Lite dễ dàng hơn so với các bản khác.OWL DL (OWL Description Logic): hỗ trợ cho những người dùng cần sự diễn cảm tối đa trong khi cần duy trình tính tính toán toàn vẹn (tất cả các kết luận phải được đảm bảo để tính toán) và tính quyết định (tất cả các tính toán sẽ kết thúc trong khoảng thời gian hạn chế). OWL DL bao gồm tất cả các cấu trúc của ngôn ngữ OWL, nhưng chúng chỉ có thể được sử dụng với những hạn chế nào đó (Ví dụ: Trong khi một lớp có thể là một lớp con của rất nhiều lớp, một lớp không thể là một thể hiện của một lớp khác).
OWL DL cũng được chỉ định theo sự tương ứng với logic mô tả, một lĩnh vực nghiên cứu trong logic đã tạo nên sự thiết lập chính thức của OWL.OWL Full muốn đề cập tới những người dùng cần sự diễn cảm tối đa và sự tự do của RDF mà không cần đảm bảo sự tính toán của các biểu thức. Ví dụ, trong OWL Full, một lớp có thể được xem xét đồng thời như là một tập của các cá thể và như là một cá thể trong chính bản thân nó. OWL Full cho phép một ontology gia cố thêm ý nghiã của các từ vựng được định nghĩa trước (RDF hoặc OWL).
Các phiên bản này tách biệt về các tiện ích khác nhau, OWL Lite là phiên bản dễ hiểu nhất và

phức tạp nhất là OWL Full.

Mối liên hệ giữa các ngôn ngữ con của OWL:

- Mọi ontology hợp lệ dựa trên OWL Lite đều là ontology hợp lệ trên OWL DL
- Mọi ontology hợp lệ dựa trên OWL DL đều là ontology hợp lệ trên OWL Full
- Mọi kết luận hợp lệ dựa trên OWL Lite đều là kết luận hợp lệ trên OWL DL
- Mọi kết luận hợp lệ dựa trên OWL DL đều là kết luận hợp lệ trên OWL Full

Công cụ để xây dựng các Ontology là **Protégé**. Công cụ này được sử dụng để tạo ra file OWL.

Quy trình xây dựng Ontology

1. Ontology learning

Ontology Learning có thể được mô tả như là việc thu thập của 1 mô hình miền từ dữ liệu (miền ở đây có thể như là: Geographical,...). Ontology learning cần dữ liệu đầu vào để học những khái niệm liên quan đến miền đã biết trước, những định nghĩa của khái niệm cũng như các mối quan hệ tổ chức giữa những định nghĩa này. Dữ liệu đầu vào có thể là lược đồ như là XML-DTD, những mô hình UML, hoặc lược đồ cơ sở dữ liệu. Ontology learning có được thực hiện trên cơ sở của các nguồn được cấu trúc như XML hoặc tài liệu HTML … Trong trường hợp ontology learning được thực hiện trên cơ sở của các nguồn văn bản không được cấu trúc, chúng ta sẽ nói về ontology learning from text.

Ontology learning from text

Ontology learning có thể xem là 1 quá trình của công nghệ đảo mã (reverse engineering). Tác giả của 1 văn bản hoặc 1 tài liệu về 1 mô hình miền trong ý thức và bắt đầu tác giả chia sẻ ý tưởng với những tác giả khác để viết những tài liệu về cùng 1 miền. Tác vụ xây dựng lại mô hình thế giới của tác giả hoặc thậm chí mô hình mà được chia sẻ bởi các tác giả khác nhau, có thể được xem như là 1 loại của công nghệ đảo mã (reverse engineering).

Tác vụ này rất phức tạp và thử thách bởi vì 2 lý do:

- Đây chỉ là 1 phần nhỏ tri thức về miền của những tác giả và quy trình của công nghệ đảo mã có thể xây dựng lại mô hình của tác giả.
- Tri thức thế giới, chúng ta đang xem xét 1 quyển sách hoặc cuốn từ điển – nó ít khi đề cập rõ ràng. Chỉ 1 phần liên quan của tri thức mà trong văn bản hoặc 1 bài cáo được đề cập nhiều hoặc ít rõ ràng.

  Tri thức thế giới được chứa đựng trong những văn bản theo cái cách mà những từ và những cấu trúc ngôn ngữ được sử dụng bởi tác giả. Điều đó gây khó khăn vì mỗi tác giả sử dụng những từ, cấu trúc ngôn ngữ của họ, và không theo 1 quy ước nào, gây khó khăn trong việc lấy dữ liệu cho việc ontology learning.

2. Phát triển ontology

Việc phát triển ontology chủ yếu liên quanđến việc tiên đề hóa (axiomatize) định nghĩa của những khái niệm (concepts) cùngvới mối quan hệ (relations) giữa chúng. Đối với 1 vài ứng dụng của ontologies,điều quan trọng là kết nối những khái niệm và quan hệ đến những kí hiệu (symbols) mà được sử dụng để tham chiếu đến chúng. Điều này nghĩa là việc thu thập tri thức ngôn ngữ học về những thuật ngữ mà được sử dụng để tham chiếu đến 1 khái niệm cụ thể và những từ đồng nghĩa có thể có của những thuật ngữ này. Sau đó, 1 ontology bao gồm cây phân cấp khái niệm, các quan hệ không phân cấp. Để ràng buộc việc giải thích của những khái niệm và quan hệ, biểu đồ tiên đề (axiom schemata) như là sự phân biệt đối với các khái niệm như symmetry, reflexivity, transitivity,... Cuối cùng, cũng là 1 trong những quan tâm đến việc sử

dụng 1 ontology để lấy được dữ kiện mà không được mô hình hóa rõ ràng trong cơ sở tri thức nhưng có thể được thu từ nó.

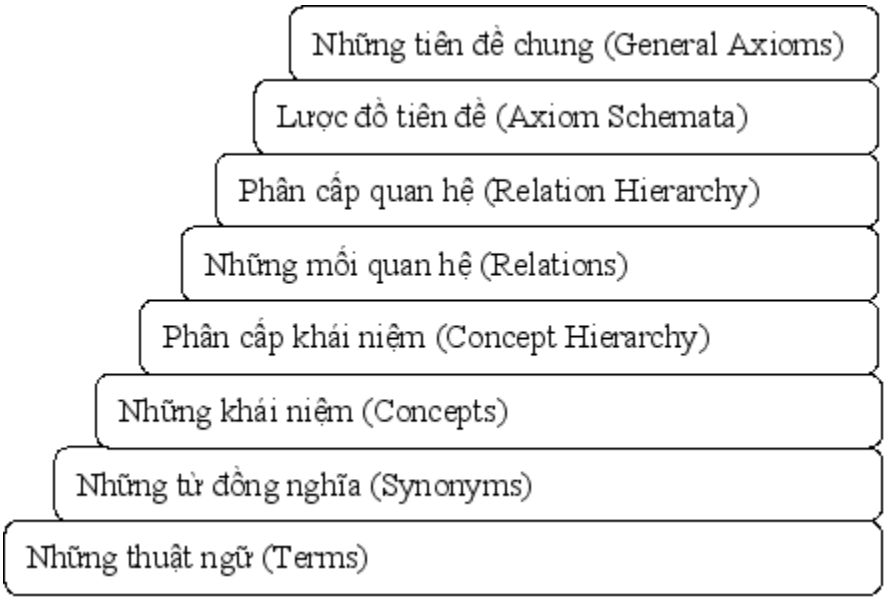
Phân lớp đưa ra những tác vụ phụ khác nhau của việc ontology learning:

- việc thu thập những thuật ngữ liên quan, ngôn ngữ
- sự nhận dạng những thuật ngữ đồng nghĩa, những biến thể
- hệ thống khái niệm (concepts),
- việc tổ chức phân cấp các khái niệm (concepts),
- và phạm vi thích hợp learning những quan hệ (relations), thuộc tính với miền
- việc tổ chức phân cấp những mối quan hệ (relations),
- instantiation of axiom schemata
- khái niệm những tiên đề tùy ý (arbitrary axioms)

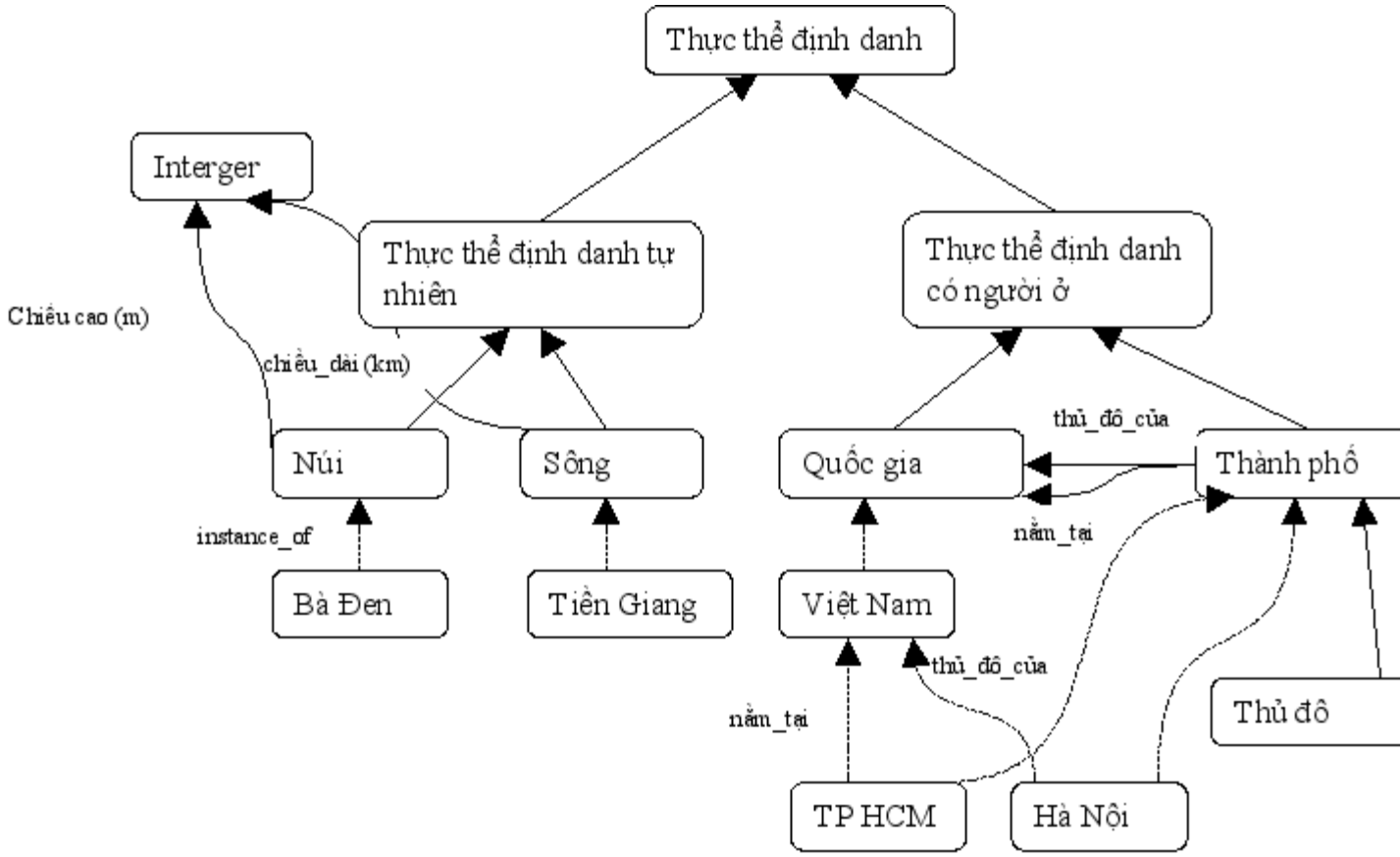
Trong hầu hết các trường hợp, những lớp xây dựng dựa trên những lớp ở phía dưới đã xây dựng rồi. Những quy trình ở những lớp cao hơn phụ thuộc vào output của những quy trình tương ứng ở các lớp thấp hơn.

Tuy nhiên, những tác vụ ở những lớp khác nhau có thể được nhóm lại với nhau và sử dụng cùng 1 thuật toán.

Ontology learning layer cake



Ở bước thu thập thuật ngữ, chúng ta sẽ tìm những thuật ngữ liên quan như sông, quốc gia, nước, thành phố, thủ đô. Tại bước tìm từ đồng nghĩa, chúng ta nhóm quốc gia và nước như là khái niệm tương đương. Tiếp theo, chúng ta learning phân cấp khái niệm giữa những khái niệm. Đối với miền địa lý, có thủ_đô $\leq_C$ thành_phố, thành_phố $\leq_C$ thực_thể_có_người_ở (Inhabited_GE).



Thêm vào nữa, chúng ta learning các mối quan hệ với nhau như là mối quan hệ thủ_đô_của giữa thành_phố và quốc_gia. Tại cấp độ biểu đồ tiên đề (axiom schemata), chúng ta thu được sông và núi là những khái niệm phân biệt. Cuối cùng, chúng ta lấy những quan hệ phức tạp hơn giữa các khái niệm và quan hệ trong hình thái tiên đề. Ví dụ: quy định nói rằng quốc_gia có 1

thủ_đô duy nhất.

3. Những tác vụ của ontology learning

3.1 Xác định thuật ngữ (Terms):

Những thuật ngữ là sự nhận dạng ngôn ngữ học của những khái niệm về lĩnh vực cụ thể. Tác vụ ở đây chính là tìm ra tập hợp những thuật ngữ hoặc dấu hiệu cho các khái niệm và quan hệ, mà chính là đặc điểm của lĩnh vực cụ thể, và sẽ cung cấp cơ sở để định nghĩa 1 bộ từ vựng (lexicon) cho ontology.

Những thuật ngữ có thể là từ đơn hoặc từ ghép mà có ý nghĩa với lĩnh vực đã cho. Đầu vào cho tác vụ này là 1 tập hợp những tài liệu liên quan đến lĩnh vực (domain) quan tâm, và đầu ra là tập hợp chuỗi S_C và S_R : chứa đựng những thuật ngữ mà được dùng như là dấu hiệu cho khái niệm và quan hệ.

Trong đó S_C là dấu hiệu cho khái niệm, S_R là dấu hiệu cho quan hệ, nằm trong khái niệm Lexicon.

3.2 Xác định từ đồng nghĩa (Synonyms):

Tác vụ khám phá từ đồng nghĩa bao gồm việc tìm những từ mà có khái niệm tương tự. Chúng ta chú ý rằng 2 từ được xem là đồng nghĩa nếu chúng có nghĩa chung mà có thể được dùng như là cơ sở để hình thành 1 khái niệm liên quan đến lĩnh vực.

Chú ý rằng có 1 sự chồng chéo giữa khái niệm đồng nghĩa và mối quan hệ từ vựng *cohyponymy*. Cohyponymy được định nghĩa là mối quan hệ giữa hyponyms và hypernym.

Ví dụ : spoon is a hyponym of cutlery

musical instrument is a hypernym of piano

3.3 Những khái niệm (Concepts):

Sự hình thành khái niệm cung cấp:

- định nghĩa của những khái niệm
- sự mở rộng của những khái niệm
- những dấu hiệu từ vựng được dùng để tham chiếu đến chúng.

Chúng ta định nghĩa 1 khái niệm gồm 3 phần < i(c), [|c|],Ref_C(c) >i(c) is the intension of the concept [|c|] : sự mở rộng của khái niệm Ref_C(c) : mô tả sự nhận dạng từ vựng trong bộ ngữ liệu (corpus)

3.4 Xác định phân cấp khái niệm (Concep Hierarchies)

Có những tác vụ liên quan :

Việc đưa vào cấu trúc phân cấp khái niệm (Concept Hierarchy Induction) :

Ví dụ : bắt đầu
từ tập khái niệm C:= { Thực thể định danh, Thực thể định danh tự nhiên, Thực thể định danh có người ở, Núi, Sông, Quốc gia, Thành phố }, công việc phải làm là sẽ đưa ra $\leq_C$ (phân cấp khái niệm hoặc phân loại tư duy (taxonomy))

Núi < thực thể định danh tự nhiên, sông < thực thể định danh tự nhiên, thực thể định danh tự nhiên < thực thể định danh, quốc gia < thực thể định có người ở, thành phố <
thực thể định có người ở, thủ đô < thành phố, thực thể định danh có người ở < thực thể định danh.