Semantic Web và thư viện số

1. World Wide Web và những hạn chế của nó

Hệ thống mạng toàn cầu đã trở nên rộng khắp thông qua một loạt các tiêu chuẩn được thiết lập rộng rãi và đảm bảo được các thành phần ở các mức độ khác nhau. Giao thức TCP/IP đảm bảo rằng chúng ta không phải lo lắng về việc chuyển từng bit dữ liệu thông qua hệ thống mạng nữa. Tương tự như vậy, HTTP (HyperText Transfer Protocol) và HTML (HyperText Markup Laguague) đã cung cấp các cách để có thể nhận thông tin và trình diễn các tài liệu siêu văn bản. Tuy nhiên, có một khối lượng không lồ các tài nguyên thông tin trên Web, điều này làm nảy sinh vấn đề là làm thế nào để tìm kiếm chính xác tài nguyên mình mong muốn. Dữ liệu trong các file HTML có thể hữu ích ở ngữ cảnh này nhưng vô nghĩa đối với ngữ cảnh khác. Ví dụ: Chúng ta biết mã vùng (***Post Code***) và muốn tìm địa chỉ của nó, nhưng mỗi quốc gia có tên hệ thống mã vùng khác biệt và Web không biểu diễn được mối liên hệ này, nên chúng ta không nhận được điều chúng ta mong đợi. Trái lại, đối với Semantic Web, chúng ta có thể chỉ ra kiểu của mối liên hệ này. Ví dụ: ***Zip Code*** (mã quốc gia) tương đương với ***Post Code*** (mã vùng). Vì vậy, nếu như các thành phần chính yếu của dữ liệu trong Web trình bày theo dạng thức thông thường, thì khó sử dụng dữ liệu này một cách phổ biến.

2. Sự ra đời của Semantic Web

Thế hệ web đầu tiên là những trang HTML thủ công, thế hệ thứ hai đã tạo nên một bước ngoặt cho máy thực hiện thường là các trang HTML động. Thế hệ web thứ ba là " Semantic Web – Web ngữ nghĩa", mang mục đích là thông tin sẽ do máy xử lý. *Semantic Web* sẽ làm cho các dịch vụ thông minh hơn. Ví dụ: Môi giới thông tin, tác nhân tìm kiếm, bộ lọc thông tin v.v. Những dịch vụ thông minh trên hệ thống web giàu ngữ nghĩa như thế chắc hẳn sẽ vượt trội hơn những phiên bản sẵn có hiện tại của các dịch vụ này.

2.1 Semantic Web là gì?

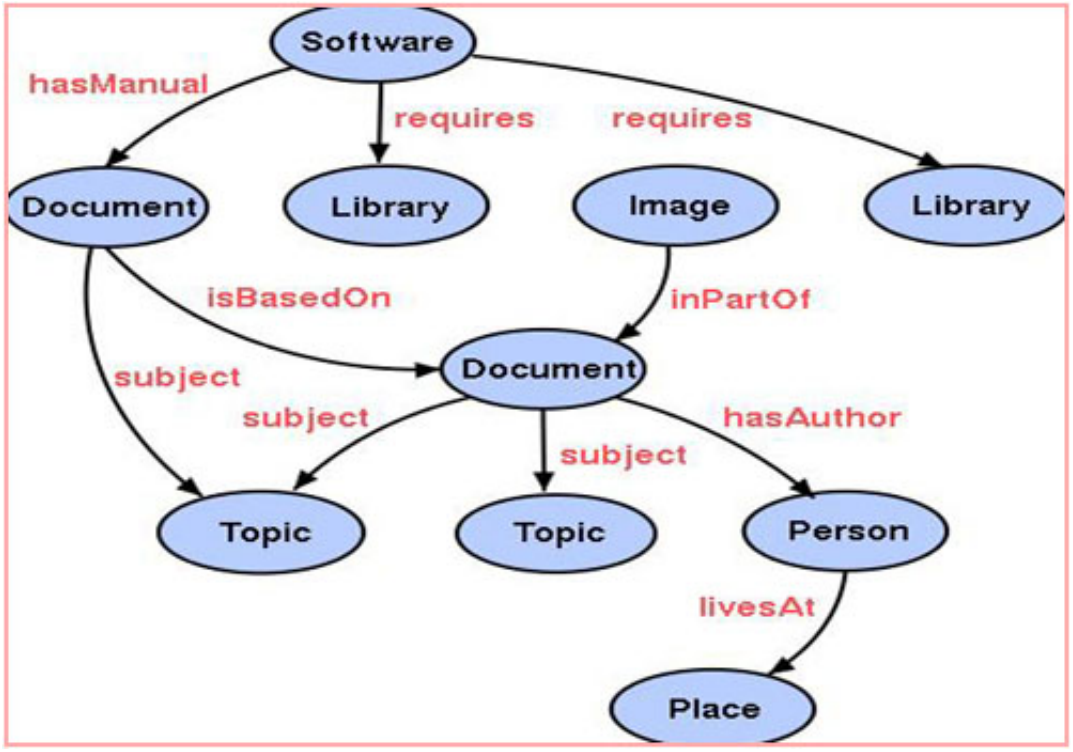
Semantic Web không là Web riêng biệt mà là một sự mở rộng của Web hiện tại, theo cách thông tin được xác định ý nghĩa tốt hơn, nó cho phép máy tính và người cộng tác với nhau tốt hơn. Semantic Web được hình thành từ ý tưởng của Tim Berners-Lee, người phát minh ra WWW (World Wide Web), URI (Uniform Resource Identification), HTTP, và HTML. Semantic Web là một mạng lưới các thông tin được liên kết sao cho chúng có thể được xử lý dễ dàng bởi các máy tính ở phạm vi toàn cầu. Nó được xem là cách mô tả thông tin rất hiệu quả trên World Wide Web, và cũng được xem là một cơ sở dữ liệu có khả năng liên kết toàn cầu. Semantic Web là một phương pháp cho phép định nghĩa và liên kết dữ liệu một cách có ngữ nghĩa hơn nhằm phục vụ cho máy tính có thể "hiểu" được. *Semantic Web* còn cung cấp một môi trường chia sẻ và xử lý dữ liệu tự động bằng máy tính.

Ví dụ: Giả sử ta cần so sánh giá để chọn mua một bó hoa hay ta cần tra cứu catalog của các hãng chế tạo xe khác nhau để tìm ra thiết bị thay thế cho các bộ phận bị hư hỏng của xe Volvo 740. Thông tin mà ta thu được trực tiếp trên Web có thể trả lời các câu hỏi này nhưng đòi hỏi con người phân tích ý nghĩa của dữ liệu và sự liên quan của nó với yêu cầu đề ra, không thể xử lý tự động bằng máy tính.

Với *Semantic Web* ta có thể giải quyết vấn đề này bằng 2 cách:

Thứ nhất: Nó sẽ mô tả chi tiết dữ liệu. Do đó một chương trình xử lý không cần quan tâm đến các định dạng (format), hình ảnh, quảng cáo trên một trang Web để tìm ra sự liên quan của thông tin.

Thứ hai: *Semantic Web* cho phép chúng ta tạo ra một file mô tả mối liên hệ giữa các tập dữ liệu khác nhau. Ví dụ: Ta có thể tạo một liên kết semantic giữa cột mã quốc gia '***zip-code***' trong cơ sở dữ liệu (database) với trường '***zip***' ở trên giao diện (form) nhập liệu nếu chúng có chung ý nghĩa. Điều này cho phép máy tính theo các đường kết nối và tích hợp dữ liệu từ nhiều nguồn khác nhau. Ý tưởng liên kết các nguồn khác nhau (tài liệu, hình ảnh, con người, khái niệm,…) cho phép chúng ta mở rộng Web thành một môi trường mới với tập các mối quan hệ mới giữa các nguồn dữ liệu, tạo ra các mối liên hệ ngữ cảnh (contextual relationship), điều mà Web hiện tại chưa làm được.



Liên kết ngữ nghĩa giữa các nguồn khác nhau trong Semantic Web

2.2. Semantic Web mang lại những gì?

2.2.1. Máy có thể hiểu được thông tin trên Web

Internet ngày nay dựa hoàn toàn vào nội dung. Web hiện hành chỉ cho con người đọc chứ không dành cho máy hiểu. Semantic Web sẽ cung cấp ý nghĩa cho máy hiểu.

Ví dụ:

The Beatles là một ban nhạc nổi tiếng của thành phố Liverpool.

John Lennon là một thành viên của The Beatles.

Bản nhạc "Hey Dude" do nhóm The Beatles trình bày.

Những câu như thế này có thể hiểu bởi con người nhưng làm sao chúng có thể được hiểu bởi máy tính? Semantic Web là tất cả những gì về cách tạo một Web mà cả người và máy có thể hiểu. Người dùng tin sẽ vẫn có thông tin trình bày theo cách trước đây, nhưng đối với máy tính, Semantic Web sẽ làm cho máy hiểu được nghĩa và tìm ra thông tin chính xác hơn Web hiện hành. Bây giờ, máy không phải suy luận dựa vào ngữ pháp và các ngôn ngữ đánh dấu (Markup Language) nữa vì cấu trúc ngữ nghĩa của văn bản (text) thực sự đã chứa nó rồi.

2.2.2. Thông tin được tìm kiếm nhanh chóng và chính xác hơn

Với *Semantic Web*, việc tìm kiếm sẽ dễ dàng nếu mọi thứ được đặt trong ngữ cảnh. Ý tưởng chính yếu là toàn bộ ngữ cảnh mà người sử dụng được biết đến. Mục tiêu của *Semantic Web* là phát triển các tiêu chuẩn và kĩ thuật để giúp máy hiểu nhiều thông tin trên Web, để máy tìm ra các thông tin dồi dào hơn, tích hợp, duyệt dữ liệu, và tự động hóa các thao tác. Với *Semantic Web*, chúng ta không những nhận được những thông tin chính xác hơn khi tìm kiếm thông tin từ máy tính, mà máy tính còn có thể tích hợp thông tin từ nhiều nguồn khác nhau, biết so sánh các thông tin với nhau.

2.2.3. Dữ liệu liên kết động

Với *Semantic Web*, chúng ta có thể kết hợp các thông tin đã được mô tả và giàu ngữ nghĩa với bất kì nguồn dữ liệu nào. Ví dụ: Bằng cách thêm các**metadata** (*siêu dữ liệu*) cho các tài liệu khi tạo ra nó, chúng ta có thể tìm kiếm các tài liệu mà metadata cho biết tác giả là ***Eric Miller***. Cũng thế, với metadata chúng ta có thể tìm

kiếm chỉ những tài liệu thuộc loại tài liệu nghiên cứu.

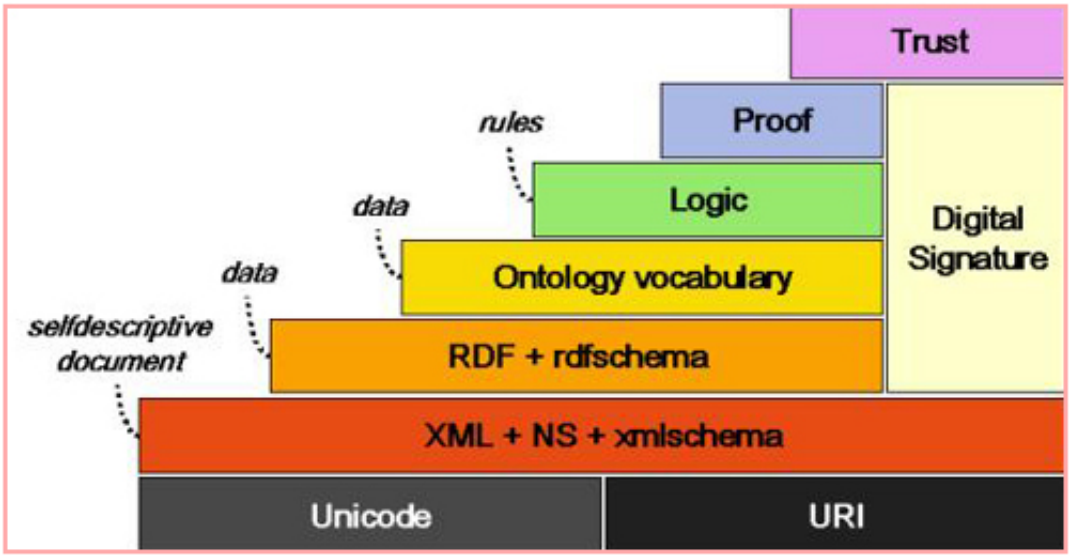
Với *Semantic Web*, chúng ta không chỉ cung cấp các URI cho tài liệu như đã làm trong quá khứ mà còn cho con người, các khái niệm, các mối liên hệ. Như trong ví dụ trên, bằng cách cung cấp những định danh duy nhất cho mỗi con người như vai trò của '***tác giả***' và khái niệm '***tài liệu nghiên cứu***', chúng ta đã làm rõ *người* ở đây là ai và mối quan hệ tương ứng của người này với một tài liệu nào đó. Ngoài ra, bằng cách làm rõ người mà chúng ta đang đề cập, chúng ta có thể phân biệt những tài liệu của ***Eric Miller*** với những tài liệu của những người khác. Chúng ta cũng có thể kết hợp những thông tin đã được mô tả ở nhiều site khác nhau để biết thêm thông tin về người này ở những ngữ cảnh khác nhau. Ví dụ như vai trò của anh ta ra sao khi anh ta là tác giả, nhà quản lý, nhà phát triển.

2.2.4. Hỗ trợ công cụ tự động hóa

Ngoài những lợi ích trên, *Semantic Web* còn cung cấp các loại dịch vụ tự động từ nhiều vùng khác nhau: từ gia đình và các thư viện kĩ thuật số cho đến các dịch vụ kinh doanh điện tử và dịch vụ sức khỏe.v.v. *Semantic Web* cung cấp phương tiện để thêm các thông tin chi tiết lên Web nhằm hỗ trợ sự tự động hóa cho các dịch vụ.

2.3 Kiến trúc Semantic Web

Semantic Web là một tập hợp/một chồng (stack) các ngôn ngữ. Tất cả các lớp của Semantic Web được sử dụng để đảm bảo độ an toàn và giá trị thông tin trở nên tốt nhất.



Kiến trúc Semantic Web

- Lớp **Unicode & URI:** Bảo đảm việc sử dụng tập kí tự quốc tế và cung cấp phương tiện nhằm định danh các đối tượng trong Semantic Web. URI đơn giản chỉ là một định danh Web giống như các chuỗi bắt đầu bằng "http" hay "ftp" mà bạn thường xuyên thấy trên mạng (ví dụ: http://www.cadkas.com).Bất kỳ ai cũng có thể tạo một URI, và có quyền sở hữu chúng. Vì vậy chúng đã hình thành nên một công nghệ nền tảng lý tưởng để xây dựng một hệ thống mạng toàn cầu thông qua đó.

- Lớp **XML** cùng với các định nghĩa về *namespace* (vùng tên gọi) và *schema* (lược đồ) bảo đảm rằng chúng ta có thể tích hợp các định nghĩa Semantic Web với các chuẩn dựa trên XML khác.

- Lớp **RDF [RDF] và RDFSchema [RDFS]:** ta có thể tạo các câu lệnh (*statement)* để mô tả các đối tượng với những từ vựng và định nghĩa của URI, và các đối tượng này có thể được tham chiếu đến bởi những từ vựng và định nghĩa của URI ở trên. Đây cũng là lớp mà chúng ta có thể gán các kiểu (*type*) cho các tài nguyên và liên kết. Và cũng là lớp quan trọng nhất trong kiến trúc Semantic Web .

- Lớp **Ontology**: hỗ trợ sự tiến hóa của từ vựng vì nó có thể định nghĩa mối liên hệ giữa các khái niệm khác nhau. Một Ontology (bản thể luận trong logic) định nghĩa một bộ từ vựng mang tính phổ biến & thông thường, nó cho phép các nhà nghiên cứu chia sẻ thông tin trong một hay nhiều lĩnh vực.

- Lớp **Digital Signature**: được dùng để xác định chủ thể của tài liệu (ví dụ: tác giả hay nhan đề của một loại tài liệu).

- Các lớp **Logic, Proof, Trust:** Lớp logic cho phép viết ra các luật (rule) trong khi lớp proof (thử nghiệm) thi hành các luật và cùng với lớp trust (chấp nhận) đánh giá nhằm quyết định nên hay không nên chấp nhận những vấn đề đã thử nghiệm.

3. Ứng dụng của semantic web

3.1. Xây dựng các bộ máy tìm tin

Vấn đề hiện nay là đa số các bộ máy tìm tin đều thực hiện cho phép người sử dụng có thể tạo các câu truy vấn gồm các từ khóa tìm kiếm để nhận về kết quả mong muốn. Tuy nhiên, phương pháp này gặp hai vấn đề chính sau đây:

• Mỗi từ khóa có thể có một hay nhiều ý nghĩa tùy theo từng ngữ cảnh và bộ máy tìm kiếm không thể hiện mối quan hệ giữa các từ khóa với nhau.

• Có thể các thông tin cùng ý nghĩa với thuật ngữ trong biểu thức tìm của người sử dụng sẽ không tồn tại trong kết quả tìm.

Ví dụ: ta cần tìm thông tin về người trưởng bộ môn công nghệ thông tin của MIT, ta gõ: "*MIT information technology chair*" vào Google, nhưng kết quả thu được là không chính xác. Nguyên nhân của việc tìm kiếm thất bại là do: Từ khoá "*MIT*" có nhiều ý nghĩa. Ngoài ra, máy tìm không thể hiểu mối liên hệ giữa các từ khoá: *MIT*, *information technology* và *chair*. Nếu bộ máy tìm kiếm được tích hợp tri thức để hiểu được ý nghĩa của các từ, thì rất có thể nó cho ta kết quả chính xác hơn, lúc đó việc tìm kiếm sẽ dựa trên khái niệm (concept) chứ không phải theo từ khóa (keyword).

3.2. Ứng dụng công nghệ ngữ nghĩa trong thư viện số:

Thư viện số phải thường xuyên xử lý một lượng lớn thông tin từ các dạng tài liệu số. Phần lớn chúng được rút ra từ thư viện truyền thống, được tập trung biên tập lại thành nguồn thông tin sẵn dùng cho một nhóm người liên quan bằng cách quét bài báo, sách, tài liệu… Bằng cách này đã làm hạn chế lợi thế của các hệ thống máy tính hiện đại và gây khó khăn cho quá trình xử lý sau này. Áp dụng công nghệ semantic web chúng ta có thể nghiên cứu và phát triển hệ thống thư viện số có thể thực hiện xử lý, lưu trữ, tìm kiếm và phân tích tất cả các kiểu thông tin số. Công nghệ ngữ nghĩa cho phép miêu tả đối tượng, thiết lập các lược đồ cần thiết trong các dạng của ontologies cho các định danh của các đối tượng số. Mục tiêu chính là làm cho thao tác giữa các phần có thể xử lý thông minh, nhất quán, mạch lạc tương tự các lớp của đối tượng số và các dịch vụ.

Ứng dụng ontologies trong việc mô tả hệ thống thư mục: Thông thường một thư viện số sử dụng dữ liệu mô tả có cấu trúc để mô tả hệ thống thư mục tuy nhiên các trường trong dữ liệu mô tả lại không được định nghĩa ngữ nghĩa một cách đầy đủ, việc ứng dụng ontologies trong thư viện số không những thực hiện lưu trữ dữ liệu mô tả để mô tả hệ thống thư mục mà còn mô tả được nội dung của nó. Thay vì trong trường hợp một quyển sách được lưu trữ trong thư viện số chúng ta có thể tách riêng cấu trúc từng chương của nó, cung cấp mô tả cho mỗi chương và thực hiện lưu trữ mối quan hệ của các chương khác nhau. Bằng việc sử dụng tư tưởng cấu trúc của ontologies và sử dụng tư tưởng này trong việc mô tả dữ liệu, chúng ta cung cấp một tầng tổng quát dữ liệu mô tả và nội dung.

Một trong những ứng dụng quan trọng nữa chúng ta có thể thấy hệ thống dữ liệu của thư viện số rất lớn và đa dạng nó thường phục vụ cho nhiều tổ chức, cá nhân vào nhiều mục đích khác nhau, trong khi đó dữ liệu chủ yếu thuộc vào hai dạng là dữ liệu có cấu trúc (trong database) và dữ liệu phi cấu trúc (các nguồn lấy từ web). Một vấn đề dăt ra là làm thế nào để các ứng dụng sử dụng được đồng thời cả hai loại dữ liệu này, bởi vì trên thực tế mỗi ứng dụng chỉ sử dụng một loại dữ liệu có cấu trúc hoặc phi cấu trúc. Chúng ta có chuẩn chung phục vụ cho hầu hết các loại ứng dụng đó là sử dụng XML (Extensible Markup Laguage), nó được xem là nền tảng công nghệ của semantic web. Nó sẽ là cầu nối thực hiện chuẩn hoá các nguồn dữ liệu, từ đó có thể phục vụ cho mọi loại ứng dụng.

3.3. Khung làm việc để quản lý tri thức (Framework for Knowledge Management)

Semantic Web là một hệ nền nhiều hứa hẹn cho việc phát triển các hệ thống quản lý tri thức. Tuy nhiên, vấn đề ở đây là làm thế nào biểu diễn tri thức ở dạng thức máy có thể hiểu được, để tri thức cần thiết có thể được tìm thấy bởi các máy tìm (search engine). Chúng ta sử dụng giải pháp quản lý tri thức dựa trên định dạng tương thích RDF để biểu diễn các luật và dựa trên một kỹ thuật mới để chú giải các nguồn tri thức bằng cách sử dụng các câu điều kiện. Giải pháp là dựa trên các công cụ Semantic Web đang tồn tại. Điểm thuận lợi chính là sự thúc đẩy khả năng tìm kiếm tri thức với độ chính xác cao, cũng như khả năng truy cập cấu tạo các nguồn tri thức cần thiết cho việc giải quyết một vấn đề nào đó. Dạng thức này có thể được biểu diễn bằng cách dùng các câu

lệnh If–Then (statement If-Then), được thiết lập theo cách suy diễn (inference) và ủy quyền (trust) trên Semantic Web. Các statement (*câu lệnh*) điều kiện có thể được dùng để lập chỉ mục nội dung các tài nguyên Web một cách nhiều ý nghĩa hơn so với liên kết các từ khóa, khái niệm hay metadata (*siêu dữ liệu*). Điều này có thể sẽ hình thành các truy vấn dựa trên ngữ cảnh hơn, tăng cường độ chính xác trong tìm kiếm tri thức. Ví dụ: Trong vấn đề định chỉ mục tài liệu, dù có hay không có tài liệu được định chỉ mục bằng từ khóa aspirin (*thuốc aspirin*) và headache (*bệnh đau đầu*), cách aspirin trị headache hay aspirin gây ra headache đều có thể được giải quyết dễ dàng bằng cách sử dụng các câu điều kiện định nghĩa trước. Việc xây dựng và quản lý tri thức trên Semantic Web một cách khoa học cho phép sự chuyển đổi đa dạng trong môi trường phân tán.

4. Kết luận:

Internet ra đời đã mang lại nhiều hữu ích cho con người, đặc biệt là trong tìm kiếm thông tin. Tuy nhiên việc tìm tin trên mạng thường bị nhiễu và nhiều khi rất khó lựa chọn được thông tin cần thiết. Semantic Web ra đời hy vọng sẽ sớm khắc phục được những nhược điểm này, góp phần nâng cao hiệu quả của mạng toàn cầu trong việc tìm và khai thác thông tin của người dùng

Tài liệu tham khảo

1. Kruk Sebastian Ryszard, Decker Stefan, Zieborak Lech. **Adding Semantic Web Technologies to Digital Libraries. -** 2005. http://library.deri.ie/
2. Nguyễn Văn Triều Dâng. **Ứng dụng web ngữ nghĩa vào phân tích trực tuyến:** Luận văn thạc sĩ CNTT. - TP. Hồ Chí Minh: Đại học Công nghệ thông tin, 2006. – 115 tr.
3. Sebastian Ryszard Kruk1, Bernhard Haslhofer, Piotr Piotrowski, Adam Westerski, Tomasz Woronieckil **- The Role of Ontologies in Semantic Digital Libraries. -** paper 2007. http://www.glam.ac.uk
4. http://www.w3.org/2001/sw/

Nguyễn Công Nhật

(Nguồn: Tạp chí Thư viện Việt Nam)