VN-KIM CHO WEB VIỆT CÓ NGỮ NGHĨA

VN-KIM FOR VIETNAMESE SEMANTIC WEB

Cao Hoàng Trụ

Khoa Công Nghệ Thông Tin, Đại học Bách khoa, Tp. Hồ Chí Minh, Việt nam

BẢN TÓM TẮT

World Wide Web đang chuyển sang một thế hệ mới là Web có ngữ nghĩa, trên đó các trang Web không chỉ có con người mới đọc và hiểu được, mà máy tính cũng có thể hiểu và xử lý chúng tự động. Bài báo này giới thiệu một dự án mà chúng tôi đang thực hiện nhằm xây dựng và phát triển một hệ thống chú thích ngữ nghĩa và khai thác thông tin trên Web tiếng Việt, tên VN-KIM. Hệ thống bao gồm một cơ sở tri thức về các thực thể có tên ở Việt Nam, như nhân vật, tổ chức, công ty, tỉnh thành, núi non, sông ngòi, và các địa điểm đặc biệt; một thành phần chú thích ngữ nghĩa tự động; và một thành phần truy hồi tài liệu có chú giải. Một số kết quả hiện tại của dự án cũng sẽ được trình bày.

ABSTRACT

World Wide Web is moving to its next generation, called Semantic Web, in which not only humans can read and understand web pages, but computers can also understand and process them automatically. This paper introduces a project that we are doing, in order to build and develop a system for semantic annotation and information exploitation on Vietnamese Web, named VN-KIM. It consists of a knowledge base about named entities in Vietnam, such as people, organizations, companies, cities, mountains, rivers, and locations of special interest; an automatic semantic annotation component; and an annotated document retrieval component. Current results of the project are also presented.

1. GIỚI THIỆU

Như chúng ta đã thấy, World Wide Web (gọi tắt là Web) đã trở thành một kho tàng thông tin khổng lồ của nhân loại và một môi trường chuyển tải thông tin không thể thiếu được trong thời đại công nghệ thông tin ngày nay. Sự phổ biến và bùng nổ thông tin trên Web cũng đặt ra một thách thức mới là làm thế nào để khai thác được thông tin trên Web một cách hiệu quả, mà cụ thể là làm sao để máy tính có thể trợ giúp xử lý tự động được chúng. Muốn vậy, trước hết máy tính phải hiểu được thông tin trên các tài liệu Web, trong khi ở thế hệ Web hiện tại thông tin được biểu diễn dưới dạng văn bản thô mà chỉ con người mới đọc hiểu được.

Điều này đã thúc đẩy sự ra đời của ý tưởng *Web có ngữ nghĩa* (Semantic Web), một thế hệ mới của Web, mà lộ trình phát triển của nó đã được Tim Berners-Lee, cha đẻ của Web, phác thảo ra chỉ khoảng sáu năm về trước ([1]). Web có ngữ nghĩa là sự mở rộng của Web hiện tại mà trong đó thông tin được định nghĩa rõ ràng sao cho con người và máy tính có thể cùng làm việc với nhau một cách hiệu quả hơn. Mục tiêu của Web có ngữ nghĩa là để phát triển các chuẩn chung và công nghệ cho phép máy tính có thể hiểu được nhiều hơn thông tin trên Web, sao cho chúng có thể hỗ trợ tốt hơn việc khám phá thông tin, tích hợp dữ liệu, và tự động hóa các công việc.

Bài báo này trình bày nội dung của một dự án mà chúng tôi đang thực hiện, để phát triển một hệ thống chú thích ngữ nghĩa và khai thác thông tin trên Web tiếng Việt, tên VN-KIM. Phần 2 tóm tắt các hướng nghiên cứu chính hiện nay về Web có ngữ nghĩa. Phần 3 trình bày việc thiết kế và hiện thực VN-KIM, với các kết quả đã đạt được. Cuối cùng, Phần 4 đưa ra một số kết luận.

2. WEB CÓ NGỮ NGHĨA

2.1. Các hướng nghiên cứu

Hiện tại, các hoạt động nghiên cứu về Web có ngữ nghĩa đang tập trung vào ba hướng chính sau đây ([7]):

- Chuẩn hoá các ngôn ngữ biểu diễn dữ liệu (XML) và siêu dữ liệu (RDF) trên Web.
- Chuẩn hoá các ngôn ngữ biểu diễn Ontology cho Web có ngữ nghĩa.
- Phát triển nâng cao Web có ngữ nghĩa (Semantic Web Advanced Development - SWAD).

Trong hướng thứ ba về SWAD nói trên, một vấn đề được các nhà khoa học quan tâm nhất và cũng là nền tảng nhất của Web có ngữ nghĩa là làm thế nào để nhúng ngữ nghĩa vào các tài liệu Web, mà hiện nay được viết bằng ngôn ngữ tự nhiên và chỉ có con người mới đọc hiểu được. Hơn nữa việc nhúng ngữ nghĩa này phải được thực hiện một cách tự động để có thể chuyển đổi hàng tỷ các tài liệu Web đã có sẵn sang các tài liệu tương ứng cho Web có ngữ nghĩa. Muốn vậy, vấn đề đầu tiên cần giải quyết là rút trích tự động ngữ nghĩa của mỗi tài liệu Web rồi chú thích lại ngữ nghĩa này vào tài liệu đó.

Trong một tài liệu, các thực thể có tên được đề cập đến tạo nên phần quan trọng cho ngữ nghĩa của tài liệu đó. Nói cách khác, để nắm được ngữ nghĩa của một tài liệu thì trước hết cần nắm được ngữ nghĩa của các thực thể có tên trong tài liệu đó.

Thực thể có tên là con người, tổ chức, nơi chốn, và những đối tượng khác được tham khảo bằng tên. Các thực thể có tên khác về bản chất và ngữ nghĩa với các *từ* (Word) ở chỗ chúng nói về các cá thể, trong khi các từ nói về những cái chung như khái niệm, phân loại, quan hệ, thuộc tính. Việc xử lý các từ do vậy chỉ đòi hỏi ngữ nghĩa từ vựng và lý lẽ thông thường, trong khi việc xử lý các thực thể có tên cần đến tri thức cụ thể về thế giới đang xem xét.

Ngữ nghĩa của các thực thể có tên tuy chỉ là một phần ngữ nghĩa của toàn bộ tài liệu, nhưng nếu có thể rút trích và chú thích chúng một cách tự động với độ chính xác tương đối cao thì cũng đã có ý nghĩa thực tiễn rất lớn. Một ứng dụng rất rõ ràng là xác định và cung cấp tự động thông tin về các thực thể có tên trong các trang Web tin tức cho người đọc.

Các tài liệu Web có chú thích ngữ nghĩa cho các thực thể có tên cũng sẽ giúp cho việc tìm kiếm và khai thác thông tin trên đó được chính xác và hiệu quả hơn. Ví dụ một truy vấn về thành phố Sài Gòn sẽ được trả về các tài liệu đề cập đến TP.HCM hoặc "Sài Gòn" như một thành phố, chứ không phải các tài liệu chứa từ "Sài Gòn" như trong "Đội bóng Cảng Sài Gòn", "Xí nghiệp may Sài Gòn", hay "Cty Saigon Tourist".

Việc xác định ngữ nghĩa cho các thực thể có tên là không đơn giản và không thể chỉ dựa vào các từ điển, vì các thực thể khác nhau có thể có cùng tên. Ví dụ để xác định xem thực thể mà từ "Sài Gòn" trong một tài liệu ám chỉ đến là một thành phố hay là một đối tượng loại nào khác, cần phải biết được ngữ cảnh nơi từ đó xuất hiện.

Vì vậy một hệ thống chú thích ngữ nghĩa cho các thực thể có tên cần có các thành phần cơ bản sau:

- Ontology: định nghĩa các lớp thực thể, bao gồm sự phân loại của các khái niệm thực thể và quan hệ giữa chúng.
- Các danh hiệu thực thể: phân biệt các thực thể với nhau và được liên kết với các mô tả ngữ nghĩa của chúng.
- Cơ sở tri thức: mô tả các thông tin cụ thể về các thực thể.

2.2. Các ứng dụng thực tế

Một số hệ thống chú thích ngữ nghĩa cho các thực thể có tên đã và đang được phát triển, trong đó KIM (Knowledge & Information Management) của Ontotext Lab, Bulgaria, tỏ ra là một hệ thống được phát triển một cách bài bản và đạt được những kết quả đáng chú ý nhất ([8]). Miền dữ liệu mà KIM nhắm vào là các thực thể được đề cập đến trong các tin tức quốc tế hàng ngày. Ontology của KIM hiện có khoảng 250 lớp, như Object (cho các thực thể con người, nơi chốn, ...) hay Happening (cho các thực thể sự kiện, hoàn cảnh, ...), và 100 thuộc tính, như subRegionOf cho lớp Location hay hasPosition cho lớp Person. Cơ sở tri thức của KIM hiện có khoảng 80,000 thực thể về các nhân vật, thành phố, công ty, và tổ chức quan trọng và phổ biến nhất trên thế giới.

HÌNH 2.1: MỘT PLUG-IN CỦA KIM

Để lưu trữ và quản lý Ontology và tri thức, KIM sử dụng Sesame ([6]), một kiến trúc tổng quát để lưu trữ hiệu quả và truy vấn đa dạng dữ liệu biểu diễn bằng RDF và RDF Schema. KIM sử dụng công nghệ Lucene ([5]) để đánh chỉ mục và truy hồi các tài liệu XML đã được chú thích ngữ nghĩa, nhưng theo các thực thể có tên thay vì theo các từ khoá.

Để rút trích thông tin về các thực thể có tên, KIM sử dụng GATE ([4]), một hệ thống gồm nhiều thành phần và công cụ để xử lý ngôn ngữ tự nhiên. Các công nghệ này đều là mã nguồn mở, đã ổn định, và được sử dụng cũng như mở rộng cho nhiều ứng dụng khác nhau.

Phiên bản hiện tại của KIM đã đạt đến độ chính xác chung về chú thích ngữ nghĩa tự động vào khoảng 80% so với con người chú thích bằng tay. Hình 2.1 minh họa một Plug-in của KIM phát hiện và nối kết tự động các thực thể có tên (ví dụ Intel) trên một trang tin điện tử của CNN vào cơ sở tri thức của chúng.

Tuy nhiên, KIM vẫn còn các hạn chế chính sau:

- Ontology dựa trên các khái niệm chung trên thế giới, trong khi ở Việt Nam có những khái niệm riêng, đặc thù cho hoàn cảnh kinh tế và xã hội của đất nước. Ví dụ, những khái niệm như "Trường bán công" hay "Thành phố trực thuộc trung ương" cần có những lớp thích hợp để xếp loại chúng.
- Động cơ rút trích ngữ nghĩa chỉ dùng cho tiếng Anh, chưa hỗ trợ được cho tiếng Việt với các đặc thù riêng về từ vựng và cấu trúc cụm từ. Ví dụ, trong tiếng Anh các từ cách nhau bằng khoảng trống, còn trong tiếng Việt một từ có thể do nhiều âm tiết cách nhau bằng khoảng trống ghép lại, như "Đất nước" hay "Công ty".
- Động cơ truy hồi thông tin chưa cho phép so trùng bán phần các truy vấn và dữ liệu lưu trữ. Các dạng truy vấn về các thực thể còn đơn giản và cứng nhắc, chưa cho phép khai thác nhiều và tiện lợi các thông tin lưu trữ trong Ontology và cơ sở tri thức đã xây dựng.

3. THIẾT KẾ VÀ HIỆN THỰC VN-KIM

3.1 Tổng quan

Trong dự án này, mục tiêu chú thích ngữ nghĩa đặt ra trong đề tài Web Việt có ngữ nghĩa được giới hạn ở mức của các thực thể có tên. Như đã nhấn mạnh trong [8], việc chú thích ngữ nghĩa này cần đến một Ontology và một cơ sở tri thức ban đầu đủ lớn, mà một trong những vai trò chính của chúng là để giải quyết sự mập mờ khi nhận dạng thực thể. Ở Việt Nam hiện chưa có một hệ thống thông tin nào quản lý tương đối đầy đủ và thống nhất các thực thể có tên.

Một cơ sở hạ tầng tri thức như vậy cho Web có ngữ nghĩa là một công việc lâu dài, cần được đầu tư xây dựng dần dần từ nhỏ đến lớn. Vì vậy, để có định hướng tập trung trong việc thu thập dữ liệu, trước mắt miền các thực thể có tên của đề tài được xác định bao gồm các nhân vật, thành phố, công ty, và tổ chức quan trọng và phổ biến nhất ở Việt Nam, cụ thể là các thực thể được đề cập đến trong các tin tức nội địa. Kết quả thu được là một Ontology và cơ sở tri thức ban đầu cho Web có ngữ nghĩa ở Việt Nam, mà sau đó có thể được bổ sung và phát triển thêm dần.

Mã nguồn mở Sesame được khai thác và mở rộng để quản lý Ontology và cơ sở tri thức nói trên dưới dạng RDF và RDF Schema. Bản thân các tài liệu Web có chú thích ngữ nghĩa được đánh chỉ mục bằng Lucene, một mã nguồn mở được áp dụng cho tiếng Việt và các thực thể có tên ở Việt Nam.

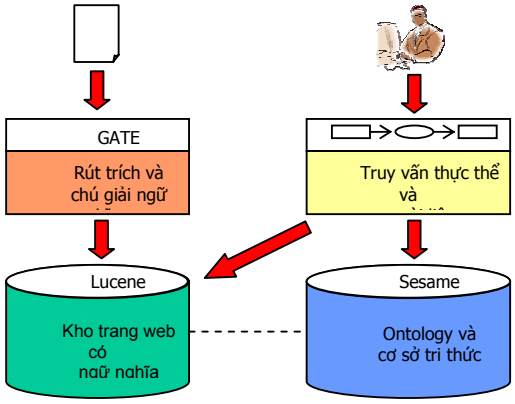
Vấn đề quan trọng còn lại là rút trích thông tin và chú thích ngữ nghĩa tự động cho các thực thể có tên trong các tài liệu Web. Để tránh các nhược điểm của phương pháp nhận dạng thực thể có tên cổ điển, hệ thống rút trích thông tin dựa trên Ontology của GATE được sử dụng. Tuy vậy, việc làm này cũng không đơn giản vì tất cả các lớp xử lý của hệ thống đều phải được

lập trình lại, dù vẫn ít tốn kém hơn nếu phải phát triển từ đầu tất cả các công cụ xử lý ngôn ngữ tự nhiên mà GATE đã có.

Để truy vấn tìm các thực thể một cách gần đúng, độ tương tự giữa các đồ thị tri thức đã được nghiên cứu. Để tận dụng động cơ tìm kiếm chính xác của Sesame và ngôn ngữ truy vấn SeRQL, phương pháp biến đổi truy vấn được áp dụng. Đồ thị ý niệm được sử dụng làm một ngôn ngữ truy vấn thân thiện và linh hoạt, bên cạnh SeRQL và các khuôn mẫu cố định.

Sau đó, các ngôn ngữ và kỹ thuật lập trình Web thông thường như JavaScript, Applet, và Plug-in được sử dụng để viết các ứng dụng và giao diện đồ hoạ cho phép chú thích tự động ngữ nghĩa của các thực thể có tên trong các tài liệu Web tiếng Việt, và truy vấn cơ sở tri thức cùng các tài liệu đã được chú thích ngữ nghĩa.

Các ứng dụng này mang đến một hình ảnh mới về người đọc tin trên Web, theo đó bên cạnh một *trình duyệt* (Browser) người đọc còn có thêm các Plug-in để truy xuất hệ thống chú thích ngữ nghĩa và cơ sở tri thức từ xa khi muốn biết thêm chi tiết về các thực thể có tên trong tin tức đang đọc. Sơ đồ tổng quát của hệ thống VN-KIM được phác thảo trong Hình 3.1.



Hình 3.1: Tổng quan về VN-KIM

3.2 Cơ sở tri thức

Tính thông minh cần đến tri thức. Riêng cho trường hợp của các hệ thống rút trích ngữ nghĩa tự động như VN-KIM, đó là tri thức về các thực thể có trên xuất hiện trên các tài liệu Web tiếng Việt. Trước khi xây dựng một cơ sở tri thức như vậy, cần phải thiết kế một lược đồ, tức một Ontology, cho nó.

KIMO là một Ontology cấp trên đơn giản nhưng đủ đáp ứng yêu cầu của việc sinh chú giải ngữ nghĩa tổng quát. Tuy nhiên, các thuộc tính và quan hệ của các lớp trong KIMO còn sơ sài. Ngoài ra, mục tiêu của xây dựng VN-KIMO là xây dựng một Ontology đặc thù cho các thực thể có tên ở Việt Nam (ví dụ "Hội Liên Hiệp Phụ Nữ", "Mặt Trận Tổ Quốc", ...). Do đó, mặc dù vẫn dựa trên KIMO, nhưng VN-KIMO có một số thay đổi chính sau:

- Các thuộc tính và quan hệ của các thực thể đầy đủ hơn để cho phép các ứng dụng có thể khai thác hiệu quả Ontology và cơ sở tri thức của hệ thống.
- Nhiều khái niệm của tiếng Anh chỉ tương ứng với một khái niệm tiếng Việt, và ngược lại. Ví dụ, "Bay" hay "Gulf" trong tiếng Anh cùng tương ứng với "Vịnh" trong tiếng Việt, còn "Vị Trí" hay "Nơi Chốn" trong tiếng Việt cùng tương ứng với "Location" trong tiếng Anh.
- Một số quan hệ phân cấp khái niệm được hiệu chỉnh cho phù hợp với các thực thể ở Việt Nam. Tuy nhiên, sự thay đổi này là tối thiểu để đảm bảo VN-KIMO có tính tương thích cao với KIMO. Điều này giúp cho VN-KIMO vẫn phù hợp với các thực thể nước ngoài.

Dựa trên VN-KIMO, một cơ sở tri thức về các thực thể ứng với các lớp thực thể trong VN-KIMO đã được xây dựng. Hiện tại, hệ thống chỉ giới hạn ở việc xây dựng cơ sở tri thức cho các miền thực thể phổ biến trong tin tức hàng ngày sau đây: nhân vật, tổ chức, công ty, tỉnh thành, núi non, sông ngòi, đường sá, điểm đặc biệt (thắng cảnh, di tích, ...). VN-KIMO có 373 lớp và 114 tính chất, và cơ sở tri thức của VN-KIM có trên 85,000 thực thể.

3.3 Chú thích ngữ nghĩa

Thành phần rút trích thông tin là một trong những thành phần quan trọng nhất của hệ thống VN-KIM. Giống như trong hệ thống KIM, nhiệm vụ của thành phần này là sinh chú thích cho các thực thể có tên trong tài liệu bao gồm kiểu thực thể và liên kết tới thực thể tương ứng trong cơ sở tri thức (nếu có).

Thành phần này bao gồm các thành phần con và được sắp xếp theo một trật tự nhất định tạo thành một mô hình rút trích thông tin. Trong đó, hai thành phần cơ bản là thành phần tra cứu ngữ nghĩa và thành phần văn phạm so trùng

mẫu. Thành phần văn phạm so trùng mẫu trực tiếp sinh ra các chú thích cho thực thể trong tài liệu, dựa trên các thông tin từ thành phần tra cứu ngữ nghĩa. Ví dụ, trong đoạn văn bản "... ông Nguyễn Văn A ...", nhờ vào ngữ liệu "ông" mà hệ thống nhận ra được "Nguyễn Văn A" là một con người.

Hỗ trợ cho thành phần văn phạm so trùng mẫu là thành phần gán từ loại tự động cho các văn bản cần chú thích. Cụ thể, trong một mẫu so trùng có thể qui định các từ ở những vị trí nhất định thuộc về những từ loại nhất định. Ví dụ, một luật so trùng mẫu nói rằng nếu một thực thể xuất hiện sau các giới từ như "ở" hoặc "tại" thì thuộc về lớp nơi chốn hoặc tổ chức.

Sau giai đoạn so trùng mẫu, để có thể tạo các liên kết thực thể vào cơ sở tri thức, cần phải giải quyết sự đồng tham chiếu và mập mờ định danh. Hai tên khác nhau gọi là đồng tham chiếu nếu chúng cùng chỉ đến một thực thể. Còn mập mờ định danh là trường hợp một tên nhưng có thể chỉ đến các thực thể khác nhau. Ví dụ trong đoạn văn bản "... TP.HCM ... thành phố Sài Gòn ...", hai tên "TP.HCM" và "Sài Gòn" là đồng tham chiếu. Một ví dụ về sự mập mờ định danh là "đường Trần Hưng Đạo" trong đoạn văn bản có thể chỉ đến một con đường ở Hà Nội hoặc con đường cùng tên ở Sàigòn.

3.4 Truy hồi thông tin

Một trong những dịch vụ tiềm năng được xây dựng dựa trên nền tảng VN-KIM đó là dịch vụ truy hồi tài liệu dựa trên thực thể có tên. Đây là một phương pháp truy hồi thông tin mới trong đó nhu cầu tìm kiếm thông tin của người dùng không chỉ được biểu diễn bằng những từ khóa mà còn được biểu diễn bằng những thực thể có tên được đề cập đến trong tài liệu. Nhu cầu tìm kiếm thông tin của người dùng được biểu diễn tốt hơn, và do đó hiệu quả của quá trình tìm kiếm thông tin cũng cao hơn.

Mô hình truy hồi thông tin theo thực thể có tên được chia thành hai khối chức năng chính là chức năng đánh chỉ mục tài liệu và chức năng tìm kiếm tài liệu theo thực thể có tên. Để truy vấn, VN-KIM cung cấp ba hình thức: (1) Mẫu hộp văn bản, đơn giản nhưng cứng nhắc; (2) SeRQL, có khả năng diễn đạt cao nhưng cú pháp phức tạp đối với người sử dụng đầu cuối; (3) Đồ thị ý niệm, dung hoà ưu và nhược điểm của hai hình thức trước ([2]).

Trên thực tế, với khối lượng thông tin khổng lồ trên Web như hiện nay, việc luôn đòi hỏi các câu trả lời chính xác là không hợp lý. Vì vậy một vấn đề đặt ra là nghiên cứu các độ đo và phương pháp truy hồi thông tin gần đúng. Hiện tại VN-KIM cho phép truy hồi thông tin gần đúng dựa trên sự tương tự và bao phủ giữa các lớp và tên thực thể ([3]).

4. KẾT LUẬN

Chúng tôi vừa giới thiệu về thế hệ sắp tới của Web là Web có ngữ nghĩa, và việc phát triển VN-KIM, một hệ thống chú thích ngữ nghĩa tự động cho các trang Web tiếng Việt. Hệ thống xác định tự động lớp của một thực thể có tên trong một trang Web, và nối kết nó với cơ sở tri thức tương ứng. Thông tin về lớp của các thực thể sẽ được nhúng vào các trang Web để cập đến chúng, làm cơ sở cho việc truy hồi tài liệu theo ngữ nghĩa hơn là từ khoá.

Hiện tại chúng tôi đang hoàn thiện cơ sở tri thức và công cụ phần mềm của VN-KIM. Việc cài đặt VN-KIM trên hệ thống máy chủ tính toán song song, và kiểm định hiệu quả của nó cũng đang được tiến hành.

Sự thừa nhận

Công trình được thực hiện dưới sự hỗ trợ kinh phí của đề tài cấp nhà nước KC.01.21 "Nghiên cứu các kỹ thuật xây dựng và khai thác thông tin Web có ngữ nghĩa". Xin cám ơn các thành viên tham gia đề tài đã giúp tác giả viết báo cáo này.

TÀI LIỆU THAM KHẢO

1. Berners-Lee T. *Semantic web roadmap*. Bản thảo đang viết, 9/1998.
2. Cao Hoàng Trụ, Đỗ Thanh Hải, Phạm Trần Ngọc Bảo, Huỳnh Ngọc Tuyên, Vũ Quang Duy. *Conceptual graphs for knowledge querying in VN-KIM*. Trong kỷ yếu của hội nghị quốc tế lần thứ 13 về "Conceptual Structures", 18-22/7/2005, Kassel, Đức.
3. Cao Hoàng Trụ, Huỳnh Tấn Đạt. *Approximate retrieval of knowledge graphs*. Trong kỷ yếu của hội nghị lần thứ 11 của "International Fuzzy Systems

Association", 28-31/7/2005, Bắc Kinh, Trung Quốc.

4. Cunningham H. et al. *Developing language processing components with GATE*. Tài liệu hướng dẫn sử dụng GATE version 2.1, 2/2003.
5. Gospodnetic O. Parsing, indexing, and searching XML with Digester and Lucene. Trong tạp chí "IBM developerWorks", 6/2003.
6. Kampman A., Harmelen F., Broekstra J. *Sesame: a generic architecture for storing and querying RDF and RDF schema*. Trong kỷ yếu của hội nghị quốc tế lần thứ 1 về "Semantic Web", 9-12/6/2002, Sardinia, Ý.
7. Koivunen M.-R., Miller E. *W3C semantic web activity*. Trong kỷ yếu của hội thảo "Semantic Web Kick-off", 02/11/2001, Phần Lan.
8. Popov, B. et al. *KIM – Semantic annotation platform*. Trong kỷ yếu của hội nghị quốc tế lần thứ 2 về "Semantic Web", 20-23/10/2003, Florida, Mỹ.