MATHIS – HỆ THỐNG HỖ TRỢ TẠO CHÚ THÍCH VÀ TÌM KIẾM TÀI LIỆU KHOA HỌC

MATHIS – SUPPORT SYSTEM FOR ANNOTATION MAKING AND SEARCH ENGINE ON SCIENCE DOCUMENTS

Lê Thành Nhân
Đại học Nice Sophia – Antipolis

Võ Trung Hùng, Cao Xuân Tuấn, Hoàng Thị Mỹ Lệ
Đại học Đà Nẵng

TÓM TẮT

Trong bài báo này chúng tôi giới thiệu những kết quả nghiên cứu bước đầu trong dự án Mathis. Đây là dự án hợp tác nghiên cứu giữa Nhóm nghiên cứu KEWI của Đại học Nice – Sophia Antipolis và Trung tâm DATIC của Trường Đại học Bách khoa – Đại học Đà Nẵng. Mục tiêu chính của dự án là nhằm biểu diễn, quản lý và tìm kiếm các công thức toán học trên môi trường web. Nội dung của dự án bao gồm việc nghiên cứu đề xuất các mô hình phù hợp, phát triển các bộ công cụ để soạn thảo công thức, soạn thảo chú thích và tìm kiếm các công thức toán học trên các tài liệu khoa học, đặc biệt là trên môi trường web. Chúng tôi đã đề xuất mô hình tổng quát cho hệ thống Mathis. Hệ thống này hoạt động dựa trên nền tảng các ứng dụng được hỗ trợ bởi tổ chức W3S và các ứng dụng phát triển trong dự án gồm bộ quản lý các chú thích, kho dữ liệu các văn bản khoa học và một bộ tìm kiếm (Search Engine).

ABSTRACT

In this paper, we present initial research results in the project Mathis (Mathematic Information web Services). This is a collaborative research project between researchers of KEWI Team (Laboratoire d'Informatique, Signaux, et Systèmes de Sophia-Antipolis (I3S) / Equipe KEWI, Université de Nice Sophia-Antipolis) and DATIC (Danang Applied of Technology of Information and Communication Center) and Danang University of Technology). The main objective of the project is to perform, manage and search for mathematical formulas on the web environment. The content of the project involves research to propose a suitable model and the development of tools for formula and annotation editing, mathematical formulas searching on the science documents, especially on the web environment. We have proposed a general model for the Mathis system. This system works on the basis of the applications supported by the W3S Organization and the applications developed from this project include tools for management of annotations, databases of scientific documents and a search engine system.

1. Giới thiệu

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của mạng Internet và công nghệ Web là sự bùng nổ thông tin số. Số lượng người sử dụng và lượng thông tin sản sinh ra trên mạng Internet gia tăng rất nhanh và chúng ta có thể tìm thấy mọi thông tin cần thiết khi có nhu cầu. Đặc biệt, lượng thông tin liên quan đến khoa học, phục vụ học tập, nghiên cứu cũng gia tăng nhanh chóng và phong phú về lĩnh vực.

Theo thống kê của Internet World Stats (http://www.internetworldstats.com/stats.htm) thì số lượng và tỉ lệ gia tăng người sử dụng Internet trong những năm qua như sau:

Khu vực	Số người dùng Internet		% dân số dùng Internet	Mức tăng 2000-2009	Tỉ lệ
	12/2000	12/2009			
Châu Phi	4,514,400	86,217,900	8.7 %	1,809.8 %	4.8 %
Châu Á	114,304,000	764,435,900	20.1 %	568.8 %	42.4 %
Châu Âu	105,096,093	425,773,571	53.0 %	305.1 %	23.6 %
Trung Đông	3,284,800	58,309,546	28.8 %	1,675.1 %	3.2 %
Bắc Mỹ	108,096,800	259,561,000	76.2 %	140.1 %	14.4 %
Mỹ La-tinh	18,068,919	186,922,050	31.9 %	934.5 %	10.4 %
Châu Đại Dương	7,620,480	21,110,490	60.8 %	177.0 %	1.2 %
Tổng cộng	**360,985,492**	**1,802,330,457**	**26.6 %**	**399.3 %**	**100.0 %**

Việc khai thác hiệu quả các tài liệu khoa học trên Web có ý nghĩa quan trọng trong khoa học và kinh tế vì nó góp phần đáng kể vào việc cải thiện quá trình học tập và nghiên cứu. Theo số liệu thống kê, khi thực hiện học tập và nghiên cứu thì con người đã chi phí khoảng 90% thời gian cho công tác tìm kiếm, phân tích và tổng hợp các tài liệu hiện có.

Hiện nay, đã có nhiều công cụ cho phép soạn thảo và quản lý các công thức toán học nhưng việc tìm kiếm nó còn gặp nhiều khó khăn. Để tìm kiếm một công thức toán học, chúng ta cần có một cơ chế thống nhất để mô tả, lưu trữ và tìm kiếm theo ngữ nghĩa tương ứng với công thức đó.

Trong báo cáo này, chúng tôi giới thiệu những kết quả nghiên cứu trong khuôn khổ dự án hợp tác giữa Trung tâm Ứng dụng Công nghệ Thông tin và Truyền thông (DATIC) của Đại học Đà Nẵng với Nhóm Nghiên cứu KEWI thuộc Trung tâm I3S (Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis) – Cộng hòa Pháp liên quan đến việc đặc tả và tìm kiếm các công thức toán học trong các tài liệu khoa học trên môi trường Internet.

2. Dự án hợp tác nghiên cứu MATHIS

Ý tưởng của dự án là đề xuất một mô hình phù hợp với các tiêu chuẩn hiện hành và cho phép chúng ta có thể mô hình hóa, lưu trữ và tìm kiếm thuận lợi các công thức toán trên các tài liệu, đặc biệt là trên các tài liệu khoa học.

Cơ sở của dự án dựa trên nền tảng đã có để hỗ trợ quản lý các công thức toán học trên môi trường web gồm:

- **MathML:** một tiêu chuẩn dựa trên nền tảng XML để quản lý các công thức toán học được đề xuất bởi W3C. MathML (viết tắt của Mathematical Markup Language - Ngôn ngữ Đánh dấu Toán học) cho phép đặc tả và cách thức lưu trữ cả cấu trúc lẫn nội dung của các công thức toán học. Ngoài ra, MathML còn

cung cấp phương thức trao đổi thông tin toán học trên máy tính (để hiển thị cũng như để tính toán) và hiển thị các tài liệu toán học trên World Wide Web [3].

- **OpenMath:** là một ngôn ngữ đánh dấu để mô tả ý nghĩa của công thức toán học. Mặt khác, nó có thể được sử dụng để bổ sung cho MathML, một tiêu chuẩn chủ yếu tập trung vào trình bày nội dung của các công thức, nhằm bổ sung thông tin ngữ nghĩa của công thức. OpenMath có thể được mã hóa trong XML hoặc trong một định dạng nhị phân. OpenMath đã được phát triển nhằm phục vụ cho việc trao đổi dữ liệu giữa các ứng dụng khác nhau [1].
- **MathNotes**: nhằm mục đích cải thiện môi trường hợp tác trên web bằng cách tạo các chú thích trong tài liệu và lưu trữ kèm theo công thức toán học, hoặc trên máy cục bộ hoặc trên một máy chủ. Môi trường được cung cấp bởi Annotea là môi trường tiêu chuẩn và có khả năng mở rộng theo yêu cầu người dùng, và được thiết kế để tích hợp với các tiêu chuẩn W3C khác. Annotea chủ yếu dựa trên các ngôn ngữ RDF/RDFS trong đó sử dụng các mẫu đại diện cho các chú thích và XPointer để liên kết các phần của một chú thích trong tài liệu [4].

Mục đích của dự án là kết nối các hướng nghiên cứu trên và phát triển bài toán theo một số hướng mới, cụ thể là:

- Xây dựng một ontology tham khảo cho các lĩnh vực toán học, hình ảnh của UMLS (Unified Medical Language System), cho lĩnh vực y tế và GO (Gene Ontology) nhằm đặc tả và lưu trữ các bộ gen. Ontology này sẽ tích hợp vào ứng dụng như các bộ sưu tập truyền thống có chứa các công thức toán (như Sổ tay Toán Chức năng của Abramowitz).
- Nhúng các đối tượng OpenMath trong một mô hình ngữ nghĩa chung để cho phép đưa vào tài khoản ngữ nghĩa của các công thức khoa học trong các chú thích và xác định các nguồn tài liệu khoa học.
- Xây dựng và thực hiện các công cụ như soạn thảo và chú thích cho các công thức, công cụ tìm kiếm ngữ nghĩa cho các tài liệu toán học (bài báo, hướng dẫn,...). Hỗ trợ cho cả việc soạn thảo văn bản và biểu diễn ngữ nghĩa của công thức.

Mục tiêu dự án Mathis, trước hết là phát triển các công cụ hỗ trợ và phần mềm để soạn thảo các chú thích và tìm kiếm theo ngữ nghĩa trên các tài liệu toán học/khoa học trên môi trường web và tiếp đến là thử nghiệm những công cụ này trong các lĩnh vực E-learning và các tài liệu khoa học.

Mathis sẽ kế thừa các kết quả nghiên cứu từ các dự án W3S, đặc biệt là các kết quả nghiên cứu về ngữ nghĩa của công thức trong các chú thích và tài liệu khoa học. Những công cụ này sẽ được thiết kế đa ngôn ngữ (Việt, Pháp, Anh) và hỗ trợ cho nhiều hệ điều hành/ngôn ngữ lập trình để phù hợp với các loại người dùng khác nhau.

3. Nội dung đề xuất

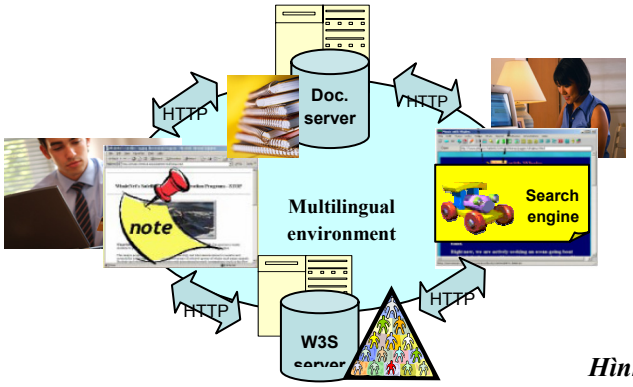
Dự án Mathis nhắm đến việc phát triển một bộ công cụ để chú thích và tìm kiếm các tài liệu khoa học trên web. Chúng tôi đề xuất 3 bộ công cụ chính sau:

- *Bộ soạn thảo công thức Mathis:* một trình soạn thảo công thức toán học để tạo ra và xuất bản các công thức toán học trực tuyến, được gọi là "eFormula" được phát triển trên cơ sở của OpenMath/MathML, kết hợp với phong cách trình bày theo kiểu web. Các mẫu công thức này được gắn kèm chú thích nhằm mô tả ngữ nghĩa liên quan đến lĩnh vực toán học.
- *Bộ tạo chú thích Mathis:* tạo ra một chú thích đính kèm công thức hoặc tài liệu khoa học. Công cụ này cho phép chúng ta tạo các đối tượng gọi là "eNote", nó sẽ được lưu trữ trên máy tính cục bộ hoặc trên một máy chủ chia sẻ. Một eNote có thể được tạo ra từ một bộ soạn thảo eNoter.
- *Bộ tìm kiếm Mathis:* Một công cụ tìm kiếm khai thác ngữ nghĩa của eNote được lưu trữ cục bộ hoặc trên một máy chủ, một cơ chế suy luận cụ thể để cung cấp câu trả lời cho người dùng [5]. Việc tích hợp bộ tìm kiếm Mathis vào các công cụ tìm kiếm trên thị trường như Google, Yahoo,... có thể sẽ được xem xét đến trong quá trình phát triển.

Những phát triển này sẽ được thực hiện sau khi nghiên cứu đề xuất một số mô hình và thực hiện một số nghiên cứu lý thuyết trong bối cảnh của dự án W3S, cụ thể:

- Định nghĩa một mô hình chính thức cho eFormula dựa chủ yếu vào OpenMath và MathML;
- Nghiên cứu về một mô hình tổng quát để mô tả mối quan hệ giữa một đối tượng toán học eFormula và khái niệm toán học trong một ontology được xây dựng trong ngôn ngữ OWL (Web Ontology Language) [2]. Mô hình này nên dẫn đến một cơ chế biểu diễn và lý luận lai (hybrid logic) để biên soạn hai loại ngữ nghĩa: ngữ nghĩa của lĩnh vực toán học và ngữ nghĩa của công thức;
- Cuối cùng, cần tiến hành các nghiên cứu về khả năng tích hợp và vai trò của việc lưu trữ và tìm kiếm các công thức toán trong các ứng dụng khác như eLearning (đào tạo trực tuyến) và eCollaboration (môi trường cộng tác trên mạng) [6].

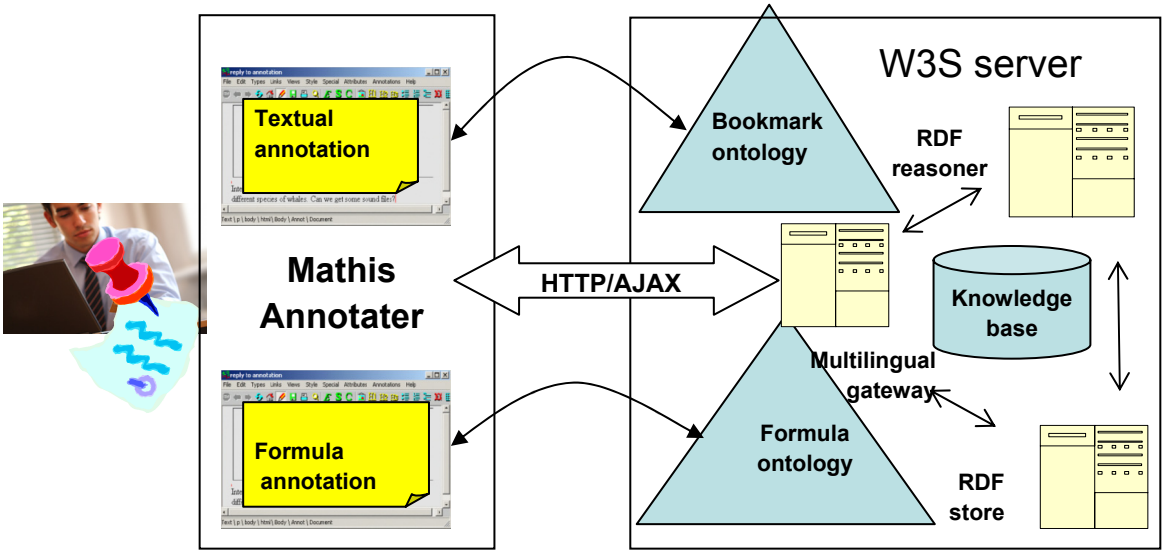
Để giải quyết các vấn đề trên, chúng tôi đề xuất trước hết một mô hình tổng quát của hệ thống như sau:



Hình 1. *Kiến trúc tổng quát của Mathis*

Hệ thống này hoạt động dựa trên nền tảng các ứng dụng được hỗ trợ bởi tổ chức W3S và các ứng dụng phát triển trong dự án gồm bộ quản lý các chú thích, kho dữ liệu các văn bản khoa học và một bộ tìm kiếm (Search Engine). Tất cả các ứng dụng này đều lưu trữ trên các máy chủ của hệ thống và tương tác với nhau dựa trên nền tảng HTTP. Hệ thống làm việc trên môi trường Internet và đa ngữ [7].

Để phục vụ cho việc tìm kiếm các công thức, chúng tôi đề xuất mô hình khởi tạo và quản lý các chú thích như sau:



Hình 2. *Hệ thống chú thích Mathis*

Hệ thống chú thích trước hết được tạo ra bởi người sử dụng thông qua bộ soạn thảo chú thích và tổ chức lưu trữ đính kèm theo công thức. Cả chú thích và công thức được lưu trữ dưới dạng ontology [4] và có thể được xử lý, tìm kiếm thông qua các máy chủ của W3S.

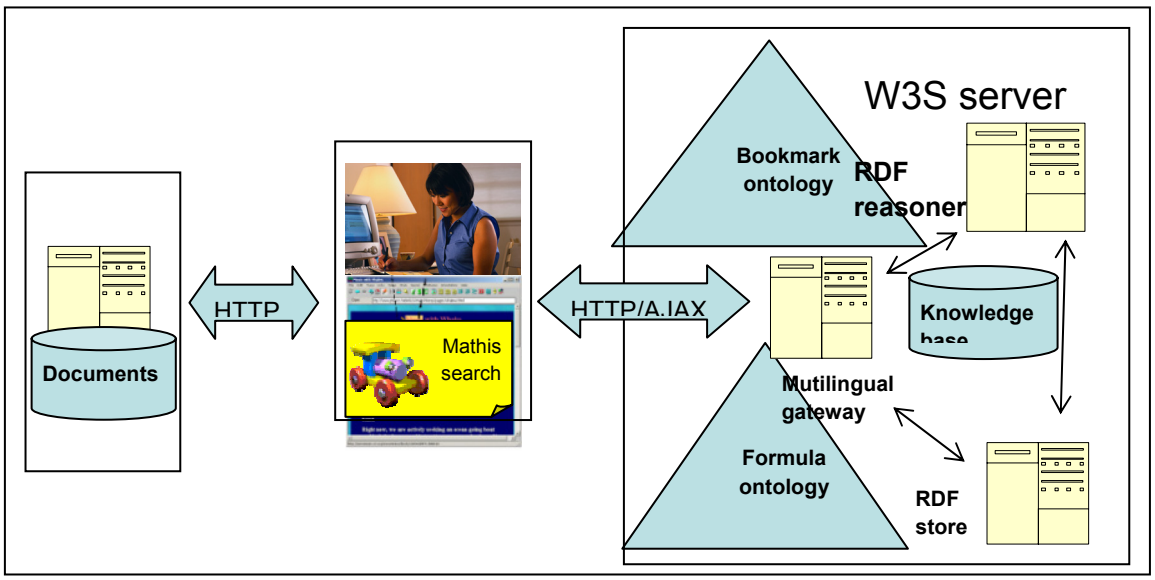


Figure 3. *Mathis Search Engine*

Để phục vụ cho việc tìm kiếm các công thức toán học, trước hết ta phải có các máy chủ có lưu trữ các tài liệu khoa học chứa các công thức toán lưu trữ theo định dạng qui ước của Mathis và được hỗ trợ bởi các máy chủ dịch vụ W3S, và cuối cùng chúng ta phát triển một bộ tìm kiếm (Mathis search) để phục vụ tìm kiếm tài liệu theo yêu cầu người dùng.

4. Kết luận

Nhu cầu tìm kiếm các công thức toán học trên môi trường web là rất lớn nhưng hiện nay chưa có hệ thống nào đáp ứng, kể cả các nhà cung cấp dịch vụ nổi tiếng như google, yahoo, Microsoft,… Việc nghiên cứu các giải pháp để hỗ trợ soạn thảo, lưu trữ và tìm kiếm các công thức toán học trên môi trường web là rất cần thiết và có ý nghĩ cao cả về mặt học thuật lẫn thực tiễn.

Đại học Đà Nẵng và Đại học Nice – Sophia Antipolis đã thành lập NiceCampus nhằm thúc đẩy các hoạt động hợp tác giữa 2 Đại học, hai bên đã phối hợp tổ chức đào tạo ở trình độ thạc sĩ cho 5 chuyên ngành (Khoa học máy tính, Quản lý nguồn nước, Hệ thống nhúng và Điện tử, Quản trị kinh doanh và E-Tourique) và triển khai một số dự án hợp tác nghiên cứu chung. Dự án Mathis là một trong những dự án nghiên cứu chung nằm trong khuôn khổ hợp tác này và góp phần tăng cường hợp tác giữa các nhà khoa học, trao đổi nghiên cứu sinh và nâng cao chất lượng đào tạo thạc sĩ cho ngành Khoa học máy tính.

Đối với dự án Mathis, hai bên đã bước đầu đề xuất được mô hình tổng quan của hệ thống và bắt đầu bắt tay vào nghiên cứu các đặc tả và phát triển các công cụ. Theo kế hoạch dự kiến, sẽ có 2 nghiên cứu sinh và 6 học viên cao học tham gia thực hiện dự án này.

TÀI LIỆU THAM KHẢO

[1] U. Buswell, O. Caprotti, D.P. Carlisle, M.C. Dewar, M. Gaëtano and M. Kohlhase: *The OpenMath Standard v 2.0*, The OpenMath Society, June, 2004.

[2] O. Corby, R. Dieng-Kuntz, C. Faron-Zucker, et F. Gandon: *Searching the Semantic Web : Approximate Query Processing based on Ontologies*, IEEE Intelligent Systems Journal, 21(1), 2006.

[3] U. Genièvre, Y. Litaiz, W. Machocki, L. Maurillon,B. Roger, S. Vallée, P. Attar: *Specification : MathML, Mathematical Markup Language*, 2003.

[4] M. Koivunen: *Annotea and Semantic Web Supported Collaboration*, Ph.D. thesis, W3C, June, 2003.

[5] A. Yurchyshyna, C. Faron, N. Le Thanh, C. Lima.: *Towards an Ontology-based approach for the compliance checking modeling in construction*, 24th W78 Int. Conference on Bringing ITC knowledge to work, Maribor, Slovenia, June 26 - 29, 2007.

[6] C. Le-Duc, N. Le-Thanh, et M. Rousset : *Compact Representation for Least Common Subsumer in Description Logic ALE*, The European Journal on Articial Intelligence - AICOM, Volume 19, Number 3, 2006, pp. 239 - 273.

[7] H. Vo-Trung, C. Boitet : *GetAMsg, une librairie pour le traitement de messages avec variantes et leur localisation dans les logiciels multilingues*, Proceeding CIDE-8, Beyrut, Lebanon, p.p. 205-222, 2005.